Report No. NAWCADWAR-94139-60

CURRENT ISSUES IN THE MEASUREMENT OF MILITARY AIRCREW PERFORMANCE: A CONSIDERATION OF THE RELATIONSHIP BETWEEN AVAILABLE METRICS AND OPERATIONAL CONCERNS

Lt. Meghan A. Carmody, U.S.N.
Air Vehicle and Crew Systems Technology Department (Code 6021)
NAVAL AIR WARFARE CENTER
AIRCRAFT DIVISION WARMINSTER
P.O. Box 5152
Warminster, PA 18974-0591

1 MARCH 1994

FINAL REPORT
Period Covering October 1993 to December 1993

Approved for Pubic Release; Distribution is Unlimited.

Prepared for
Air Vehicle and Crew Systems Technology Department (Code 6021)
NAVAL AIR WARFARE CENTER
AIRCRAFT DIVISION WARMINSTER
P.O. Box 5152
Warminster, PA 18974-0591

NOTICES

REPORT NUMBERING SYSTEM — The numbering of technical project reports issued by the Naval Air Warfare Center, Aircraft Division, Warminster is arranged for specific identification purposes. Each number consists of the Center acronym, the calendar year in which the number was assigned, the sequence number of the report within the specific calendar year, and the official 2-digit correspondence code of the Functional Department responsible for the report. For example: Report No. NAWCADWAR-92001-60 indicates the first Center report for the year 1992 and prepared by the Air Vehicle and Crew Systems Technology Department. The numerical codes are as follows:

CODE	OFFICE OR DEPARTMENT
00	Commanding Officer, NAWCADWAR
01	Technical Director, NAWCADWAR
05	Computer Department
10	AntiSubmarine Warfare Systems Department
20	Tactical Air Systems Department
30	Warfare Systems Analysis Department
50	Mission Avionics Technology Department
60	Air Vehicle & Crew Systems Technology Department
70	Systems & Software Technology Department
80	Engineering Support Group
90	Test & Evaluation Group

PRODUCT ENDORSEMENT — The discussion or instructions concerning commercial products herein do not constitute an endorsement by the Government nor do they convey or imply the license or right to use such products.

Reviewed By: Joseph E. Kern Date: 11/30/94
Branch Head

Reviewed By: [signature] Date: 11/30/94
Division Head

Reviewed By: Thomas Milhous Date: 12/8/94
Director/Deputy Director

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
1 MARCH 1994

3. REPORT TYPE AND DATES COVERED
FINAL 10/93 – 12/93

4. TITLE AND SUBTITLE
CURRENT ISSUES IN THE MEASUREMENT OF MILITARY AIRCREW PERFORMANCE: A CONSIDERATION OF THE RELATIONSHIP BETWEEN AVAILABLE METRICS AND OPERATIONAL CONCERNS

5. FUNDING NUMBERS

6. AUTHOR(S)
LT. MEGHAN A. CARMONDY, USN

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Vehicle and Crew Systems Technology Department (Code 6021)
NAVAL AIR WARFARE CENTER; AIRCRAFT DIVISION WARMINSTER
P.O. Box 5152
Warminster, PA 18974-0591

8. PERFORMING ORGANIZATION REPORT NUMBER
NAWCADWAR-94139-60

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Vehicle and Crew Systems Technology Department (Code 6021)
NAVAL AIR WARFARE CENTER; AIRCRAFT DIVISION WARMINSTER
P.O. Box 5152
Warminster, PA 18974-0591

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report discusses the primary categories of metrics used by DoD researchers to assess human performance in the aviation domain. In addition to outlining the categories and several representative metrics, various approaches are evaluated on the basis of pre-defined measurement criteria as well as the specific considerations and limitations of the particular testing environment.

14. SUBJECT TERMS
HUMAN PERFORMANCE, METRICS CRITERIA; AVIATION RESEARCH, PILOT PERFORMANCE, WORKLOAD, SITUATIONAL AWARENESS

15. NUMBER OF PAGES
62

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Current Issues in the Measurement of Military Aircrew Performance: A Consideration of the Relationship Between Available Metrics and Operational Concerns

Meghan A. Carmody
Air Vehicle and Crew Systems Technology Department
Human Factors Branch
Naval Air Warfare Center, Aircraft Division, Warminster, PA

Table of Contents

INTRODUCTION

The measurement of aircrew performance has long been an integral part of aviation history. Early research focused primarily upon issues concerning aircrew selection and training, physiological and perceptual effects, and cockpit standardization (Davis, 1948; Drew, 1940; Fitts and Jones, 1947; Wiener and Nagel, 1988; Williams, 1947). Forerunners in this arena formed, on the basis of their talents and carefully constructed research, a knowledge base establishing an inarguable link between the performance of the human element, and the performance outcome of the total system. Researchers noted, for example, that fatigued aviators neglected peripheral activities, such as monitoring fuel status, which sometimes resulted in a decline of total system performance (Davis, 1948; Wiener and Nagel, 1988). Pilots flying in poor weather conditions, or following unusual attitudes, would often experience vertigo, sometimes leading to accidents in the early history of flight, when the effects of the aviation environment on human physiology were poorly understood (Gillingham and Wolfe, 1985; Wiener and Nagel, 1988; Sanders and McCormick, 1987).

Some may point to the above as evidence that the human element is the weakest link in the aviation system chain. This may, at times, be true. However, the human element may also be argued to be the most flexible, so researchers are better suited to retain him and to try to understand him; what affects him, how he operates, and how he affects and is affected by the aircraft system. This is the goal of aircrew performance measurement. It is a goal that has proven through its history to have many challenges.

The first of these challenges involves defining the area of interest. What is the measurement of aircrew performance? The answer to this question provides an initial glance at some of the remaining challenges confronting this area of investigation.

According to Lane (1986), "the 'measurement of performance' involves the simultaneous consideration of philosophical, physical and behavioral issues" (p. 16). Lane (1986) argues that the distinctions between these three important elements defining performance measurement are critical to understanding the relationship between human and system performance, and that they are too often forgotten by researchers.

Of the three elements in Lane's (1986) definition, researchers often focus on the physical measures. Lane describes physical measurement as the measurement of mere properties and he indicates that such measurement has no direct meaning in and of itself. "In some efforts, it can be seen that performance is treated as a 'construct' for which the physical measures that can be extracted from the system are 'candidate' variables for use in assessment; 'meaning' is attached to a measure by empirically linking it to other variables or factors. This idea of 'performance as a construct' is a crucial one in measurement systems" (Lane, 1986, p. 9). When the investigator adds restrictions or tolerance limits to these values, he/she arrives at evaluative measures with a bit more meaning, such as root mean square error, targets correctly identified, etc. Although such evaluative measures give more meaning than the physical measures alone, one still does not have a concrete picture of human performance. Specifically, the roots of the human behavior have not yet been exclusively examined. Furthermore, Lane (1986) indicates that researchers frequently mistake human performance for system performance or at least fail to make a clear delineation between the two. While at the point of establishing tolerance limits to physical values, and achieving such measures as root mean square error, the researcher has still not dissected the root of the error. Is it in the human's performance, is it in the aircraft's performance, is it some combination thereof.

Experts have argued that the only way to truly understand the potential for human performance to affect system outcome is to examine the process of human performance itself (Lane, 1986; Wilson, 1991). While a great deal of emphasis is placed on terminal or outcome performance in human-machine systems, particularly in aviation research, such measures can obscure or even conceal the part in the overall process where diminishing performance has occurred (Lane, 1986). This is not to say that outcome measures should not be examined, simply that when used in isolation or without a great deal of analytical preprossessing, they are not meaningful indicators of the process of human performance. The fact is, there are a number of paths that can be taken to achieve the same outcome (Cooper and Harper, 1969; Lane, 1986; Wilson, 1991). According to Lane (1986), "identical terminal outcomes on a task can be produced by quite different orderings of procedures which represent widely divergent skill levels and energy investments. This...remains a critical weakness in the use of 'outcome' measures of proficiency" (p. 10). The key phrase in Lane's quotation is "energy expenditure".

The pilot/human element is a valuable, but quite limited, source of information processing and mechanical energy. It should be a fundamental goal of research and design to utilize such energy wisely. This means measures should concentrate heavily upon the economy of energy expenditure in the "path", because some paths are clearly superior in this respect. The fact that a more negative pathway may lead to successful overall task or mission performance under certain circumstances does not mean that it will under others. This is particularly true when comparing ideal circumstances to worst-case scenarios, as, for example, when generalizing lab results to combat performance. If one measures, in the lab, a loss of energy capacity (i.e. mental or physical energy) in a limited-capacity model like the human operator, that should be taken very seriously, regardless of whether there is an overall change in performance outcome.

To illustrate the importance of considering what occurs in the pathway of human performance, Lane (1986) discusses the "natural pilot model" of Krendal and Bloom (1963). The natural pilot model includes a characterization of a very good pilot. This characterization includes economy of effort, consistency and adaptability. Economy of effort refers to the fact that the expert pilot utilizes fewer attentional resources to maximize performance. This is similar to experts in virtually any domain, as would be indicated in the theories of automatic processing (Schneider and Shiffrin, 1977), rule versus skill versus knowledge based behavior (Rasmussen, 1986), and the ability of experts to process information in "chunks" (Chase and Simon, 1973). Consistency refers to the reliably high performance output of the expert pilot, across a variety of conditions. Finally, adaptability refers to the expert pilot's ability to effectively deal with a variety of situations.

Lane's (1986) reasoning for drawing his reader's attention to the work of Krendal and Bloom (1963) is "a key point...is that observation of outcome variables or even intermediate criterion measures are insensitive to all these factors. Maintaining control of an aircraft within 'tolerance bands' on a maneuver, for example, could result from hundreds of control inputs by a novice or only a few from a highly skilled pilot" (p. 11).

In the modern, high-tech environment, one of the primary goals in evaluating a new system or a new element within a system is the ultimate change in mission effectiveness. Human performance measures are often only of interest if they can be directly linked to

mission effectiveness, and this is typically through outcome measures. We must not fail, however, to recognize the extreme importance of examining changes in the process of human performance, as well. The human element is a crucial factor in the overall aircraft system. If any technological innovation is added to the system, a fundamental goal in testing the new device should be to assure that it enhances, or at least maintains, the pilot's ability to behave as an expert, which includes the ability to effectively manipulate limited cognitive resources. As is the case with any other vital part of the aircraft system, designers should strive to conserve the pilot's reserve energy. When a pilot's resource capacity is challenged, he can be expected to behave more erratically, like a novice. Such behavior is characterized by Lane (1986) as "creating less 'reserve' for handling unanticipated events" (p. 12). Lane goes on to suggest that "rather than focusing on the surface manifestations of 'how well the job was performed,' measurement systems should look for indicators of the three factors [of the natural pilot model] to obtain generalizable measures of true proficiency" (p. 12).

Because the researcher of human performance is attempting to measure phenomena "beneath the surface", great care is required to obtain controlled and meaningful results. As such, there are several criteria which have been established as important goals in the accurate measurement of human performance. These criteria include reliability, validity, sensitivity, and diagnosticity (Eggemeier, Biers, Wickens, Andre, Vreuls, Billman and Schueren, 1990; Lane, 1986). There are more practical than critical considerations, as well. Lane (1986) adds utility and value to the list, while Eggemeier et al (1990) think it important for the researcher to consider implementation requirements and pilot acceptance. The following section defines and describes these suggested criteria.

Human Performance Measurement Criteria

Reliability

Reliability refers to the consistency of a measure across time and like conditions, and is typically spoken of in terms of test-retest reliability (Defense and Civil Institute of Environmental Medicine [DCIEM], 1988; Fracker, 1991b). Test-retest reliability refers to the capability of a measure to provide the same results when the exact conditions are run on two

or more separate occasions. In addition to test-retest reliability, Lane (1986) also describes two other major forms of reliability: internal consistency and equivalent or alternate forms. Lane (1986) indicates that all three forms of reliability (test-retest, internal consistency and equivalent [alternate]) "estimate the same quantity.....defined as the ratio of the 'true score' variance to the total variance" (p. 50), but that all forms of reliability are derived from classical research, and may lose clarity in an operational setting such as modern aviation. Furthermore, Lane (1986) suggests that in the case of measuring highly complex tasks, there is evidence that indicates a single measure is not likely to be reliable. Therefore, in the case of such tasks as are found in modern aviation, the human performance researcher should choose a battery of metrics on which to base an assessment, and he/she should run several trials before such assessments are made.

Validity

Validity refers to the ability of a metric to actually measure what it was intended to measure (Fracker, 1991b). Lane (1986) describes validity as the process of ascribing accurate meaning or interpretation to a measured outcome.

Sensitivity

According to DCIEM (1988), the sensitivity of a measurement should be such that it is capable of distinguishing between several conditions of interest imposed on an operator. For example, if the researcher is interested in examining workload within a particular aviation situation, the sensitivity of the technique employed to assess workload would increase with the technique's capacity to measure workload variations of specific aviator tasks (Eggemeier et al., 1990).

Diagnosticity

The criterion of diagnosticity is usually discussed with respect to workload. Diagnosticity refers to the ability of a metric to determine the exact cause of an observed behavior or to discriminate between the various human information processes that may be affected by a particular manipulator. In the case of workload, for example, certain metrics are diagnostic of very specific information processing resources, such as central versus motor processing, while

other metrics are very sensitive to global workload levels, but are unable to break down the processing site of the increased attentional load (DCIEM, 1988).

DCIEM (1988) suggests that, when examining human information processing, a researcher may choose to balance sensitivity and diagnosticity by employing a battery of tests. Specifically, "a metric with high sensitivity could detect problematic situations; then a more diagnostic measure could be used to isolate the resources involved" (p. 11).

Intrusiveness

Intrusiveness refers to the degree to which a metric interferes with the normal activities of the examined environment. A very intrusive measure might, for example, interfere with a pilot's flying task or its mere presence may impose additional stress loads. The responsible researcher should be careful to select metrics which neither interfere with the operator's routine tasks, nor produce additional work by their presence. Such a consideration will help protect both the safety of the operator (especially in realistic situations, such as flight) and the clarity of the experimental results (Eggemeier et al., 1991; DCIEM, 1988).

Implementation Requirements

Before embarking upon any research endeavor, the investigator must carefully understand and plan for implementation requirements. Implementation requirements refer to any "hardware, software, training, data analysis, and other equipment or procedures necessary" (DCIEM, 1988, p. 11) to successfully complete a project. Such implementation requirements are limited by the practical constraints of the experiment and should not be so overwhelming that they become intrusive to data collection. Furthermore, the "classic" human factors considerations for any device or control modification (Wierwille and Williges, 1978) apply in the case of experimentation as well, particularly when dealing with aviation systems. Issues concerning physical space requirements, portability of equipment, data transmission and integration of the equipment into the human machine system can all be vital to collecting valid data (DCIEM, 1988).

Operator Acceptance

Eggemeier et al. (1984) point out that, particularly in aviation research, the level of operator acceptance of empirical devices may seriously affect performance outcome. Specifically, "assessment procedures that are perceived by operators as bothersome or artificial incur the risk of being ignored, or performed at substandard levels" (p. C11)

Applicability

Finally, DCIEM (1988) has added the requirement of applicability to their suggested list. They outline two criteria which are important in establishing the applicability of a metric: "the ability of the measure to reproduce in the field the results obtained in the lab...[and]... the ability of the measure to produce valid results over a wide range of loading situations" (p. 11).

The importance of these measurement criteria becomes obvious when considering the major areas involved in the measurement of human performance.

Before going into these specific areas, let us consider the rôle of the human operator in the aviation system. Ultimately, the aviator is the governing information processor. He takes in information, processes it (integrates it, makes sense of it, etc) and he responds, typically with some observable output. Undoubtedly, it is the middle stage, that of the internal processing, that presents the greatest challenge to those attempting to measure human performance, as it cannot be readily observed. However, there are numerous theoretical domains which attempt to understand all aspects of human information processing, from input to output, by examining the elements which affect the various stages.

Popular research in recent times has focused upon the internal cognitive model, composed primarily of learned associations based on training and experience, as the guide to human information processing. The information that is selected from a saturated environment, as well as the manner in which such information is processed, has been argued to be largely determined by the human's internal representation of his operating environment (Braune and Trollip, 1982; Carmody, 1993; Chase and Simon, 1973; Chechile, Eggleston, and Sasseville, 1989; Minsky, 1975).

An additional consideration in human performance research, particularly in stressful

situations, is the manner in which the general state of the human operator (both physical and mental) affects his overall information processing, in terms of input, internal processing, and output.

Therefore, in the measurement of human performance, one can consider three general areas: general operator state, the manner in which relevant information is taken in and processed by the human operator, and the observable output of the human operator.

In considering the first major area of human performance measurement, general operator state, there are two dimensions which potentially affect human performance: the physical state of the operator and the mental state of the operator. The physical state of the operator refers to such elements as energy expended by the operator, operator fatigue, and physical stress level. Such measurement is typically associated with experimental questions of environmental impact, such as the heat stress added by the aviator's wearing of Chemical-Biological-Radiation (CBR) protective gear. There are a variety of measures in existence to assess the physiological condition of the human operator. Some of the more commonly applied measures in the field of aviation research include body temperature (Iampietro, Melton, Higgins, Vaughan, Hoffman, Funkhouser, and Saldivar, 1972), heart rate (ECG) (Eggemeier et al., 1990; Iampietro, et al., 1972; Wilson et al.), and metabolic analyses (perspiration, urinary catecholamines, respiratory rate, blood analysis, etc.) (Iampietro, et. al, 1972; Soliday and Schohan, 1964).

In addition to the general physical state of the operator, the general mental state is also of interest; in particular, whether or not the operator is mentally stressed or, as is more commonly referred, mentally overloaded.

MENTAL WORKLOAD

Mental workload has been a major concern in aviation research and development for many years. Despite the vast array of work done specifically in the area of mental workload, a concise and pervasive definition can be difficult to attain. However, there is a fair amount of agreement among current experts in the field that workload is a multidimensional concept involving an interaction between the organism, its task, and the environment. In particular,

"mental workload has been characterized as being a phenomenon of task input, operator internal state, response output, or any combination of these" (DCIEM, 1988, p. 3). Upon noticing the similarity between this description of workload and that of the stages of general human information processing, the reader may note that the measure of workload essentially deals with the amount and character of the processing involved and how it is affected by the demands of the task and of the environment. Reinforcing the importance of the task and environment interaction, Eggemeier et al. (1990) note the idiosyncratic nature of mental workload. "The workload of a mission cannot be thought of as independent of the mission, the system, or the particular individuals engaged in the operation" (p. 12).

Because of the multidimensional nature of workload, as well as the fact that the phenomenon is not directly observable, researchers utilize empirically derived indicators of changes in the operator's mental load. According to DCIEM (1988), such indicators fall into three categories: task demands (time available, difficulty, etc.), operator-centered variables (motivation, effort, etc.), and performance criteria (criticality, probability of failure, etc.) and at least one measure from each category should be used to create a sensitive and diagnositic battery.

There are three major categories of workload metrics: analytical, behavioral and physiological (DCIEM, 1988).

Analytic Measures

Analytic measurement procedures apply a very systematic approach to the assessment of task demands with respect to either effort or time. They are represented by a variety of mathematical models of workload, but often involve a time-line analysis of the overall task situation (DCIEM, 1988).

A time-line analysis dissects a task into the fundamental operations and then associates each operation with specific lengths of time for completion (DCIEM, 1988). Time-line analyses are based on the assumption that when the time needed to complete a task operation is greater than the time available to complete it, the human operator will experience task overload (Wierwille and Williges, 1978). According to Wickens (1984), "time is often the

crucial parameter in predicting the performance of complex man-machine systems, and likely should not be excluded from any performance evaluation." (p. 328) However, DCIEM (1988) point out a weakness in the general time-line methodology. They argue that the fundamental operations are assumed to be performed serially, a situation not representative of aviation. There are, however, similar methods which address this problem. One such method indicated by DCIEM (1988) is Wingert's (1973) Function Interlace Technique (FIT). Not unlike the theoretical proposals of Wicken's, the FIT technique holds that certain tasks may be simultaneously performed, depending on the type of input (visual, auditory and kinaesthetic), output (motor response, vocal response and no response) and task characteristics involved. Unfortunately, the FIT technique also has its drawbacks. As DCIEM (1988) indicates, the FIT technique is presently unable to "attach accurate interlacing coefficients to the various task component pairs or to the tasks themselves" (p. 18).

Behavioral Measures

Behavioral measures include primary task measures, spare mental capacity measures and subjective opinion.

Primary Task Measures

Primary task measures directly measure operator overt responses. This is, in fact, one of the advantages of this metric; its objectivity. Additionally, the metric has the advantage of non-intrusiveness and inherent operator acceptance. Potential disadvantages of this metric call into question issues of reliability, sensitivity, and diagnosticity, as interpretation of the primary task metric is based on the assumption that under adverse workload conditions (either underloading or overloading), primary task performance will diminish (Wierwille and Williges, 1978). However, as highlighted previously, "the human operator's ability to compensate with consistent system outputs for varied task demands and environmental conditions, often renders any performance measure useless as a workload indicator" (DCIEM, 1988, p. 12). On the other hand, in cases of overload, performance output clearly diminishes (Norman and Bobrow, 1975). Therefore, according to Eggemeier et al. (1990), primary task measures have the greatest utility in such overload situations. Specifically, "primary task

measures exhibit their greatest sensitivity to variations in workload when the total task demand in a situation exceeds the pilot's capability to process information. Under non-overload situations, however, primary task performance can be invariant with increases in task demand" (p.4-8). Generally, in the case of aviation tasks, there is sufficient workload to utilize primary task measures. However, this same complexity adds a dimension of difficulty to the use of primary task metrics in the aviation environment. According to Eggemeier et al. (1990), "...while traditional applications of primary task workload measurement involve assessment of the speed and accuracy of operator responses, the capability to assess accuracy in this application is very limited, due to the complexity and multi-task nature of the pilot environment" (p. 4-13). As such, Eggemeier et al. (1990) suggest that investigators specify a limited range of acceptable parameters for pilot responses, in order to better understand and interpret any findings.

Spare Mental Capacity

A great deal of the research conducted in the field of mental workload has examined spare mental capacity (DCIEM, 1988; Wierwille and Williges, 1978). The measurement of spare mental capacity in relation to mental workload is based upon the assumption that the human operator is a limited capacity information processor (DCIEM, 1988).

There are two main categories of spare mental capacity metrics: secondary task measures (DCIEM, 1988; Eggemeier et al., 1990) and occlusion (selective blocking of visual information) (DCIEM, 1988). The present paper discusses only secondary task measures, as they are the more commonly applied.

Secondary Task Measures

Eggemeier et al. (1990) explain that one of the chief reasons for using secondary task methodology is to measure the "potential for overload" (p. 4-8). Again, they argue that primary task measures are relatively insensitive below the threshold of pilot overload. Secondary task measures have the advantage of demonstrating increases in workload that may otherwise be missed in a primary task measure. Again, this is due to the fact that the pilot/subject may be compensating for increased workload by reallocating his attention in

favor of the primary task, while neglecting the secondary task. Further advantages to utilizing the secondary task methodology include lack of intrusiveness and general operator acceptance, provided the experiment is designed in such a manner that the secondary task is not, itself, a source of added workload.

According to Eggemeier et al. (1990), the secondary task methodology has proven successful in several laboratory investigations, which generally apply one primary and one secondary task to the subject. This evidence, however, highlights the fact that secondary task metrics are often too oriented towards pure laboratory concerns to be effective in aviation research. Additionally, the great many tasks the pilot must concurrently perform in a typical aviation scenario offer a variety of potential primary tasks to be examined, so that the consideration of secondary tasks would be superfluous (Eggemeier et al., 1990).

Expert researchers in the field of mental workload assessment have suggested the use of secondary task methodology can provide useful additional information, but should not be used as a substitute for primary task measurement (DCIEM, 1988; Rolfe, 1971).

Embedded Secondary Task Procedure

The embedded secondary task procedure uses some routine task within a multi-task environment as a secondary task. According to Eggemeier et al. (1990), the embedded task methodology may prove applicable to operational environments, and there is evidence that radio communications have been successfully utilized as secondary tasks in research involving low-fidelity simulation (Shingledecker, Crabtree, Simons, Courtright and O'Donnell, 1980; Shingledecker and Crabtree, 1982).

Embedded Concurrent Task Procedure

The embedded concurrent task procedure is similar to the embedded secondary task procedure in that it is applicable in complex, multi-task environments such as aviation. Prior research has indicated that in the aviation task environment, the priority assigned by pilots to particular tasks varies across mission segments and individuals. In such cases, the embedded concurrent task procedure offers a method of assessing workload of various tasks without regard to the subjectively assigned priority of that task, via multivariate analyses.

Additionally, using the embedded concurrent task procedure, the experimenter may choose to selectively load tasks which are of particular interest to the investigation or design evaluation at hand (Eggemeier et al., 1990).

With both the embedded secondary and embedded concurrent task procedures, the experimenter faces similar advantages and disadvantages as in the application of primary and secondary task metrics. However, both embedded task procedures present the aviation researcher the additional advantage of a specific orientation towards complex, multi-task environments.

Subjective Measures

Subjective measures include a variety of techniques which require subjects to rate or otherwise indicate their perceived level of workload, often on several dimensions. As such, many researchers who consider true workload as bounded by subjective perception consider that subjective techniques have the highest face validity (Reid, Shingledecker and Eggemeier, 1981; Skipper, Rieger and Wierwille, 1986). Likewise, Johannsen, Moray, Pew, Rasmussen, Sanders and Wickens have argued that "if the person feels loaded and effortful, he is loaded and effortful, whatever the behavioral and performance measures may show" (p. 105).

Although the most commonly applied form of subjective assessment is the rating scale, other applications include interviews and questionnaires. While interviews and questionnaires may undoubtedly provide valuable information concerning subjective workload, rating scales typically offer the extraction of information in a more controlled and detailed fashion (DCIEM, 1988; Williges and Wierwille, 1979). Further support for the use of quantitative scaling techniques comes from classic research on psychometric scaling through the process of magnitude estimation (Pepermans and Corlett, 1983; Stevens, 1962). Such evidence indicates the capability of humans to apply quantitative judgements to subjective sensory experiences.

Additional advantages to subjective workload assessments relate to the criteria of sensitivity and non-intrusiveness. According to several authors, subjective ratings have shown to be fairly sensitive to workload ranges below the point of overload, unlike performance-based measures (Casali and Wierwille, 1984; DCIEM, 1988; Eggemeier, 1984). Likewise,

Wickens (1984) indicates that the advantage of subjective estimates over primary task performance measures is that subjective estimates are more sensitive to the number of activities in competition, while primary task measures are more sensitive to single task difficulty. Furthermore, because subjective assessments are typically recorded after, and sometimes prior to, a task condition, they are considered to be very non-intrusive (DCIEM, 1988; Reid, Shingledecker and Eggemeier, 1981; Skipper et al., 1986).

Despite the advantages, subjective measures are also linked with several disadvantages. Foremost of these relates to the question of validity. It can be difficult for the researcher to fully ascertain exactly what is being measured. The investigator must contend with problems such as individual variations in conceptual understanding of mental workload and the potential inability to distinguish various sources of workload, such as physical versus mental (DCIEM, 1988; Wierwille and Williges, 1978). Additionally, subjective data may be biased by the subject's willingness to participate and/or "please" the experimenter, or by perceived pressures of the experimental environment. Similarly, DCIEM (1988) points to Thornton's (1985) discussion of the von Restorff phenomenon, where the time at which a subject is asked to subjectively rate his/her workload has been shown to confound that rating. Specifically, it has been shown that the closer the subjective rating is to the peak workload within a task condition, the higher the perceived workload is likely to be.

In addition to potential validity problems, Eggemeier (1984) contends that subjective measures lack diagnosticity. As such, Eggemeier suggests that subjective measures be used for their sensitivity, as screening devices, but followed with more diagnostic measures such as a performance-based metric.

A brief survey of the literature within the field of aviation psychology indicates the pervasive use of primarily four subjective workload assessment techniques. These include the Modified Cooper-Harper Scale (MCH) (Wierwille and Casali, 1983), the Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker and Eggemeier, 1981), the Subjective WORkload Dominance Technique (SWORD) (Vidulich, 1989), and the NASA Task Load Index (TLX) (Hart and Staveland, 1988).

Regarding the established criteria of sensitivity, diagnosticity, etc., there seems to be little differentiation among the three common subjective workload measures. All three tests have

been demonstrated to be sensitive to global changes in workload levels (DCIEM, 1988; Eggemeier, 1990). Their utility as diagnostic measures, however, is somewhat limited. There are reports, however, that both the NASA-TLX and the SWAT provide some measure of diagnosticity through their division into the defined dimensions. Regarding implementation requirements, all three tasks have the advantage of requiring little equipment (can be given paper and pencil, or by personal computer). The NASA-TLX and MCH are more easily and quickly applied. (Eggemeier, 1990). Regarding intrusiveness, there is presently little empirical evidence available. However, all three tasks are typically given after completion of a task or battery of tasks and are therefore assumed to be relatively non-intrusive. On the other hand, there remains the question of the von Restorff phenomenon.

With regard to acceptance, all three measures are intuitively high in acceptance. SWAT and MCH, in particular, have been shown to have a very high level of acceptance by the operational aviation community (Eggemeier et al., 1990 and DCIEM, 1988).

Subjective Workload Assessment Technique (SWAT)

The Subjective Workload Assessment Technique (SWAT) is a multidimensional scaling technique. It is based on a substantial degree of agreement among researchers in the field that mental workload is comprised of "several factors related to task demands, operator state, and time factors" (AAMRL, 1987, p. 11). Therefore, there are three dimensions of mental workload assessed by SWAT: mental effort load, time load, and psychological stress load. Mental effort load has been defined as "the amount of attention or concentration that is required to perform a task" (AAMRL, 1987, p. 11). Time load has been defined by AAMRL (1987) as "the total amount of time available to an operator to accomplish a task as well as overlap of tasks or parts of tasks" (p. 11). Finally, psychological stress load has been defined by AAMRL (1987) as "the presence of confusion, frustration, and/or anxiety associated with task performance" (p. 11). Utilizing these dimensions, subjects assign ratings within a range of 1 - 3 (lowest to highest mental workload) for each of the dimensions, following a particular task condition. There are two distinct phases to the application of SWAT, termed scale development and event rating phases. (AAMRL, 1987; DCIEM, 1988, Eggemeier et al., 1990). The scale development phase is what most distinguishes SWAT from other subjective

workload techniques, both in terms of increased fidelity and increased difficulty in application. The scale development phase allows for individual differences, but it does involve a potentially lengthy card sorting procedure. In the scale development phase, subjects are given a deck of 27 cards. Each of the cards contains a different combination of the three levels within each of the dimensions. Subjects then sort the cards in order to reflect their rankings (from lowest to highest) in terms of the degree of subjective workload. The card sort procedure is used to test conjoint measurement axioms (AAMRL, 1987), which reflect a theory outlined by Krantz and Tversky (1971). The theory "provides for a series of axioms which, when tested with a set of data, aid in discriminating between four simple polynomial models to determine which of them best fit the set of data" (AAMRL, 1987, p. 16). Due to the potentially complex nature of such testing, there is a computer program available to conduct the analysis on-line (AAMRL, 1987). Despite the analysis program, the card-sorting procedure itself is somewhat time-consuming. Eggemeier et al. (1990) found that the original card sort procedure was too lengthy for use with the operational pilots in their investigation; the assessments simply required more time than the subjects could afford (Eggemeier et al. [1990] did utilize a modified version of the SWAT scale).

The second phase, event ratings, is relatively easy to administer. The subjects simply rate the three dimensions on the previously mentioned scale of 1 to 3. However, this phase of administration, though relatively easy to accomplish, has also been called into question by Eggemeier et al. (1990). Specifically, the authors found that, for many realistic mission scenarios, the original SWAT rating scale was insufficient to reflect the extremely high levels of workload the operational pilots were reporting.

The reliability, validity and sensitivity of SWAT have been repeatedly demonstrated in a variety of laboratory simulations involving both civilian and military aircraft and missions (DCIEM, 1988; Eggemeier et al., 1990).

An additional application of SWAT is in the domain of projective workload assessment. PRO-SWAT (Reid, Shingledecker, Hockenberger, and Quinn, 1984) requests subjects to make ratings based solely on a task description, without actually performing said task. High correlations between PRO-SWAT and actual SWAT ratings have indicated much promise for the use of this technique in projective workload assessment (Eggleston and Quinn, 1984).

NASA-Task Load Index (NASA-TLX)

The NASA-Task Load Index (TLX) is a subjective workload measure developed by the Human Performance Research Group at NASA Ames Research Center (Hart and Staveland, 1988). Like SWAT, the NASA-TLX involves a multi-dimensional scale. Subjects rate workload levels for six different subscales: Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort, and Frustration. Each subscale is rated from "Low" to "Hi" (except for own performance, which is rated from "Good" to "Poor") and an overall workload rating is computed based on a weighted average of the ratings on these subscales. Research has shown the NASA TLX to be a reliable and sensitive measure of overall workload (Vidulich, 1989; Vidulich and Tsang, 1987). In fact, some researchers have indicated that there seems to be little distinction between the NASA-TLX scaling technique and SWAT in terms of validity, and recommend the use of either as an operational tool (DCIEM, 1988; Vidulich and Tsang, 1987). In the case of the NASA-TLX, however, it does have the distinct advantage of relative ease in administration. The NASA-TLX can be fairly quickly completed by subjects using personal computers.

Cooper-Harper (CH) and Modified Cooper-Harper (MCH)

The Cooper-Harper (CH) rating scale is one of the oldest techniques for assessing subjective workload in operational aviation research. The scale was developed by Cooper and Harper in 1969 and, according to DCIEM (1988), "has been widely accepted in the aviation industry as a reliable indicator of aircraft handling qualities" (p. 26). However, DCIEM (1988) also indicate that there have been problems in applying the CH scale to workload assessments associated with other than aircraft handling problems. Therefore, in order to broaden its scope, a modified version of the CH, the Modified Cooper-Harper scale (MCH), was developed by Wierwille and Casali (1983). The MCH employs a decision-tree structure and is reported to be sensitive in a variety of workload situations (Skipper et al. 1986). DCIEM (1988) reports that the applicability of the MCH outside of assessments related to flight is controversial. However, this is somewhat irrelevant for the interests of the present paper, as it is specifically concerned with operational aviation workload assessments. Within

this boundary, "The MCH scale has a very high level of operator acceptance and it is extremely easy to implement" (DCIEM, 1988, p. 26).

Subjective WORkload Dominance (SWORD) Technique

The Subjective WORkload Dominance (SWORD) technique presents subjects with a rating sheet on which are displayed all possible pairs of tasks from a multi-task situation the subjects have just completed. The subjects are then asked to compare the tasks within each pair on the basis of which was more demanding. This raw data is used to construct a SWORD judgement matrix, from which a geometric mean can be calculated in order to determine a rating for each task. The rating for each task then "represents its workload on a ratio scale relative to all other tasks" (Vidulich, Ward, and Schueren, 1991, p. 681).

The potential benefits of SWORD have been examined primarily in the domain of human-machine interface changes in aviation systems. Results of several studies indicate that the SWORD technique is reliable and sensitive to overall workload levels between tasks. However, the SWORD technique does not examine multiple dimensions of workload. The SWORD has also shown some promise as a projective workload assessment technique, as indicated in Vidulich et al., 1991. Specifically, the researchers found high correlations between SWORD and PRO-SWORD (the projective assessment version of SWORD) ratings of HUD attitude displays by operational F-16 pilots.

Current Experimental Measures

SWORD-TLX

The SWORD-TLX is an experimental subjective workload technique currently under investigation at the Naval Air Warfare Center, Aircraft Division, Warminster (NAWCACDIVWAR). This technique combines attributes of both SWORD and the NASA-TLX so that individual tasks in a multi-task situation can be compared with each other (as in SWORD) on the basis of the dimensions set forth by the NASA-TLX. The research which has been conducted to date indicates that more information about workload associated with various tasks and task combinations can be gleaned from this combination of the SWORD and NASA-TLX measures (Carmody, Gluckman, Morrison, Hitchcock, and Warm, 1994).

However, there are some disadvantages. In particular, the investigations to date have examined the use of SWORD-TLX in rating different levels and strategies of automation in the Multi-Attribute Task (MAT) Battery (Comstock and Arnegard, 1990). This battery presents three tasks. There is some indication that the SWORD-TLX would become overly complicated, both in application and interpretation, with many more tasks. This problem may, in part, be solved with the development of an on-line presentation and analysis tool for the SWORD-TLX.

Function Allocation Simulation System (FASS)

While most subjective workload scaling techniques ask subjects for input regarding a percentage of effort they perceived to be associated with a task, there is a developing class of subjective workload techniques which focus more upon the element of time. One example of such a technique has undergone developmental testing at the NAWCACDIVWAR. This assessment technique "is based on concepts of time-constrained channel limits and time-based estimates of resource loads." (Glenn, Cohen, Wherry, Carmody, Boardway, 1993, p. 1). As such, the FASS technique may be regarded as combining analytic and subjective metrics.

Based, in part, on the theories of McCracken and Aldrich (1984) and Wickens (1984), five workload channels were defined (visual perception, auditory perception, spatial information processing, analytical information processing, and verbal information processing) as well as two response output channels (manual activity and speech). In the tradition of TAWL (Bierbaum, Fulford, and Hamilton, 1989), MAN-SEVAL (Laugery et al., 1988) and HOS (Lane, et al., 1977; 1981), an on-line workload analysis tool was developed. The Function Allocation Simulation System (FASS) steps subject matter experts (ex: operational tac-air pilots) through a tactical mission timeline. Subjects are asked to estimate the <u>time</u> demand for each resource channel, as opposed to the more commonly requested effort demand. Specifically, subjects are asked, for any particular task, what percentage of the total time to complete the task was devoted to each of the resource channels.

As this technique is still being investigated, its sensitivity, diagnosticity, etc., have yet to be determined. The authors do note that subjects may have some difficulty distinguishing between resource channels.

Physiological Correlates

For many years, researchers have been examining mental workload via physiological indicators. The use of such physiological indicators of workload is based on the assumption that when an individual's mental capacity is challenged, that individual's physiological arousal functions will activate. According to DCIEM (1988), "the definition of mental load from a physiological point of view refers to the activation of the organism, which is the operator's level of arousal or excitement" (p. 18).

There are several advantages to the use of physiological measures of mental workload. Perhaps foremost of these advantages is the fact that no overt response is needed on the part of the subject (aviator), unlike, for example, primary or secondary task measures (Eggemeier et al, 1990; Wilson, 1990). As such, the measures are collected continuously. This latter fact is an additional advantage, as the continuity of physiological measures provides the capability of recording relevant but momentary changes in workload that might otherwise be missed (Eggemeier et al., 1990).

Physiological measures are also relatively non-intrusive, but only if the subject is comfortable, both mentally and physically, with the devices (Eggemeier et al., 1990). Such comfort is usually not a problem, but the comfort level established may be quite dependent on the preparation and briefing of the experimenters. For instance, pilots, whose physiological condition is a constant factor in their career path, may be nervous at the prospect of heart monitoring during flight. It is the responsibility of the investigator to dispel any fears on the part of the aviator that his/her physical condition is somehow being scrutinized.

There are also several disadvantages to the application of physiological measures in the assessment of mental workload. Perhaps foremost of these disadvantages concerns the possibility of confounding mental and physical loading. In the words of DCIEM (1988), "in order to provide a reliable measure of mental load, the parameters measuring the activation level must be able to isolate effects resulting from mental load from those due to a host of other possible contributors" (p. 18). Additionally, there are potential disadvantages associated with implementation requirements. Physiological measurement can often involve costly and complicated equipment. Likewise, because many physiological recording techniques, such as

electro-encephalograms (EEGs) and electro-cardiograms (EKGs) involve the use of electronic equipment, there are potential problems with electrical noise. This is particularly true in realistic aviation scenarios (Eggemeier et al., 1990).

The most widely used physiological measures throughout DoD research include a variety of heart rate measures (EKGs), analyses of brain activity (EEGs), measurement of evoked cortical response (evoked EEGs), and a variety of measurements of eyeblink (Eggemeier et al., 1990; DCIEM, 1988; Wilson et al, 1982; Wilson, 1991)

Heart Rate Measures

According to Eggemeier et al. (1990), the measurement of heart rate has been the most utilized physiological assessment of workload in the aviation research community. As such, there exists a wealth of empirical data suggesting the sensitivity of this measure under a variety of circumstances in both simulated and operational environments. Increases in heart rate have been demonstrated under higher levels of workload associated with different mission segments in a variety of aircraft systems, including fighter/attack, rotary wing, and general aviation systems (Eggemeier, 1990). Furthermore, several studies have shown the capability of using heart rate measures to distinguish between pilot and weapons officer roles (Wilson, 1991), as well as lead versus wingmen in operational formation flights (Wilson, et al., 1982). In fact, the work of Wilson (1991) includes a set of physiological data collected from aircrew missions under one of the most realistic scenarios to date. Specifically, heart rate (as well as eye blink) measures were taken from bomber crew members during Green Flag exercises, which involve simulated combat in actual flight. Variations in heart rate differentiated mission segments. Furthermore, studies of F-4 crewmembers in air-to-ground training missions demonstrated the capacity for heart rate measures to distinguish between the workload levels experienced by the pilot versus the weapons system officer (WSO).

On the other hand, Eggemeier et al. (1990) note several studies where measures of heart rate have failed to distinguish between a variety of task loads, although subjective tests were sensitive to such variations. Examples of the tasks manipulated range from communications tasks and perceptual loadings (Casali and Wierwille, 1983) to variations in central processing load within the context of navigational problems (Wierwille, Rahimi, and Casali, 1985).

However, all these studies were examined within the context of simulated, rather than actual flight. One may interpret these results as indicative of a relative insensitivity of such heart rate measures to other than the very high workload levels achieved in actual flight. In fact, the work of Wilson (1991) did demonstrate that heart rate was lowest in the simulator recordings, relative to flight as a lead and flight as wing. Additionally, Wilson (1991) indicated discrepancies between his laboratory and flight findings. Although he did find increases in heart rate in the lab associated with certain tasks, Wilson indicates that "the magnitude of the increase was much larger during the flight segments than during the tracking task suggesting that the cardiac dynamics were quite different" (p. 11). This finding, in part, leads to his conclusion that "the laboratory data provided very little information about the cardiac or eye blink changes that took place during flight" (p. 13).

However, there is one advantage to the discrepancy in realism between laboratory simulations and actual flight. There are those who have argued that increased heart rates associated with certain segments of flight, such as take-off and landing, reflect emotions associated with the increased risks of these segments, rather than any significant increase in mental load. (Lindholm et al., 1987). In a simulated environment, such risks are virtually eliminated. For this reason, Lindholm and Cheatham (1983) examined the heart-rates of inexperienced subjects in computer-simulations of aircraft carrier landings. Their results were similar to those obtained in prior research involving actual flight. (Lindholm et al, 1987). Furthermore, their results prompted Lindholm et al (1987) to conduct an in-flight study which controlled for the potential emotional factors accompanying risks associated with take-off and landing flight segments. The authors reasoned that the factors contributing to increased mental workload during take-off and landing were the relatively high velocities at low altitude. This combination would increase, among other things, the pilot's attention to visual information. The investigation examined the effects of g-forces, altitude, and velocity in A-7 sorties, under three conditions of visual quality. The findings of major consequence to the present document include the fact that Lindholm and his colleagues found increasing heart rates under low altitude conditions and high velocity conditions. The results for low altitude were significant across all three conditions of visual quality, and the results for velocity were significant within one of the conditions of visual quality. Therefore, their findings supported

both laboratory and in-flight evidence suggesting that heart rate increases during flight segments associated with increased mental workload.

It appears that the potential for heart rate measures to be utilized in the lab exists, perhaps if applied in more realistic simulations. However, the sensitivity of heart rate measures is a highly relevant and important issue. Much research, due to economic constraints and safety concerns, must be examined, at least initially, in a simulated environment. Therefore, researchers must be assured of the sensitivity of instruments within this environment, as well as their generalizabilty to the operational setting.

A possible solution/compromise is suggested in Eggemeier et al. (1990). The authors indicate that a measure of heart rate variability (HRV) known as spectral analysis has shown promise as a possible measure of cognitive task difficulty. In particular, the intermediate band (.07 - .14 HZ) component of HRV has been the most frequently analyzed and is thought to be associated with the regulation of blood pressure. In relation to workload, research indicates that the power of the intermediate band decreases with corresponding increases in workload. Eggemeier et al. (1990) add that "the available evidence suggests....that the .10 Hz component may be specifically sensitive to certain cognitive aspects of task demand" (p. 4-76).

There is also evidence to suggest that different measures of heart rate should be used depending on the research questions. Eggemeier et al. (1990) have indicated that both mean heart rate and HRV have been found to correlate highly with subjective metrics, but that they differ in their sensitivity to workload levels. The evidence suggests that mean heart rate is the better measure under high workload situations, and HRV is the more sensitive measure under lower, or subthreshold, levels of workload. In either case, Eggemeier et al. (1990) point out that neither measure is particular diagnostic, although spectral analysis of the HRV, particularly in the intermediate band, may prove to be diagnostic of at least general cognitive, as opposed to more physical or perceptual, loads.

Eyeblink Activity

Eyeblink activity has been examined as an indicator of workload associated with the need for greater levels of visual attention (Eggemeier et al., 1990). The components of eyeblink

which have been examined in the flight environment include blink rate, blink duration, and blink latency.

Both blink rate and blink duration have been shown to decrease with increases in task demands (Eggemeier et al., 1990). An examination of blink rate both in simulated and actual military flight have indicated its sensitivity to pilot versus copilot roles (Stern and Skelly, 1984), as well as mission type (Wilson et al., 1982). Similarly, Wilson et al. (1991) found significantly decreased blink rates during the higher load segments of an A-7 mission scenario.

One must review such findings with caution, however, as Eggemeier et al. (1990) reports on several studies in which data on eye blink rate and duration are inconclusive. A related but alternative technique may prove a bit more promising, particularly in the domain of assessing increases in general cognitive, as opposed to strictly visual, workload. Eggemeier et al. (1990) report that eyeblink latency has been demonstrated in laboratory experiments to increase with corresponding increases in memory demands (Bauer, Goldstein, and Stern, 1987), as well as under multi-task versus single-task conditions (Sirevaag, Kramer, de Jong, and Mecklinger, 1988).

Further research is required on all three techniques of eye blink measures, particularly eyeblink duration, of which there are few studies in actual flight environments. The available evidence to date, however, indicates, as concluded by Kramer (1989) and summarized by Eggemeier et al. (1990), "blink rate may be most clearly related to the requirement to extract visual information from the environment, and...blink duration and blink latency have demonstrated more promise as measures of mental workload and task demand" (p. 4-81 - 4-82).

The use of eyeblink data has been applied under highly realistic, operational settings. For example, Anderson, Chiou and Wun (1977) examined six Army helicopter pilots in extended flight and found that blink rate increased and pupillary amplitude varied as a function of the workload associated with a flight task.

There is a related form of eye movement data collection which can be described as combining aspects of analytical, behavioral, and physiological metrics of mental load. An example is the study of Sanders, Simmons and Hofman (1979), which examined 10 UH-1H

Army helicopter aviators. The investigators recorded eye movements in terms of mean and percent of total dwell time (for a particular part of displayed information) in order to determine the amount of visual free time the navigator had available during a low-level navigation flight. The eye movement data allowed the investigators to determine that 92.2 percent of the navigator's total visual time was spent in navigation, with engine and flight instrumentation utilizing only approximately 4 percent.

Although eyeblink data may prove sensitive to global changes in workload, the utilization of these techniques in diagnosing the source of any load increases may be less fruitful, particularly if they do not involve visual information processing.

Electro-Encephalogram (EEG)

Eggemeier et al. (1990) describe two types of EEG techniques that have shown sensitivity in aviation research. Both techniques involve recordings of electrical activity through the scalp. The first involves epoch analysis, and is probably less promising than the alternative, evoked potentials recordings.

Epoch analysis typically involves the examination of four frequency bands. Two of those bands, alpha (8-13 Hz) and theta (4-7 Hz) have been shown in both simulated and actual flight to decrease in strength in correspondence with increases in flight disturbances (Eggemeier et al., 1990; Sternman, Schemmer, Dushenko, and Smith, 1987) and the difficulty of a mission segment (Eggemeier, et al., 1990; Wilson, Purvis, Skelly, Fullenkamp, and Davis, 1987).

Diagnostic capacity of epoch analysis can be considered somewhat limited. This is due in part to the fact that evidence has shown the technique to be sensitive to overall levels of arousal. It can be difficult, for example, to separate mental versus physical sources of workload (Eggemeier et al., 1990).

The evoked cortical potential can be said to involve similar analyses, but with a bit more detail than epoch analysis. "The cortical potential (EP) is that component of the EEG that represents the brain's response to a discreet stimulus" (Eggemeier et al., 1990, p. 4-86). The EP must be separated from ongoing EEG activity, and there exist several means by which to accomplish this. The distinction may be made at the analysis stage (ex: linear stepwise

discriminate function analysis and spectral analysis) or, as is more often the case, at the data-collection stage. Typically, researchers gather EPs in response to a planted stimulus, taking several samples which are later averaged with respect to time. As explained by Eggemeier et al., 1990, "this technique is possible because the evoked response to each stimulus is temporally and spatially constant, while ongoing EEG activity occurs randomly with respect to the stimulus" (p. 4-87).

The evoked cortical potential has been shown to be sensitive to changes in task difficulty in both simulated and actual flight (Eggemeier et al., 1990; Kramer, Sirevaag, and Braune, 1987; Wilson et al., 1982). Wilson et al. (1991) were able to utilize EPs to distinguish the task loads of pilots versus WSOs in actual flight. Additionally, Lewis and Rimland (1979) were able to use visual evoked potentials (VEP) to distinguish right versus left hemisphere functioning in an information processing task conducted on a large group of 28 pilots and 30 radar intercept operators (RIOs). The differences discriminated between pilot and RIO subject groups and for higher rated versus lower rated subjects within each group.

The EP component which has been studied most frequently in workload assessment is known as the P300 or P3 component. The P300 component represents the peak of the evoked response that occurs within 300-600 msecs of the presented stimulus. There is some degree of utility of the P300 component in diagnosis. There is evidence that the P300 component of the EP is affected by perceptual and central processing demands, as opposed to more physical demands (Eggemeier et al., 1990).

However, there is an important consideration in the use of either epoch or evoked cortical potential analysis techniques with respect to attaining diagnostic information in particular, and even workload levels in general. It is an unfortunate characteristic of both epoch and evoked cortical potential analyses that these measures are rather sensitive to artifacts, particularly electrical noise originating either from the surrounding environment, or from the subject him/herself (Eggemeier et al., 1990; Wilson et al, 1982). In fact, it is not uncommon that several subjects are attrited, and a relatively high percentage of trials rejected, due to interference associated with elements such as muscular movements (Eggemeier et al., 1990, Thiessen, Lay, and Stern, 1986; Wilson et al., 1982). Such artifacts can be particularly troublesome during actual flight, or during dynamic flight simulations, particularly if either

involve the onset of g-forces and the subsequent need for the crew to counter such effects with an anti-g straining maneuver.

On the other hand, both EEG techniques, and EPs in particular, show promise in aiding the understanding of brain activity under various levels of workload. They have the advantage of providing an objective, physiological look into the proverbial "black box", and their utility should be pursued. However, the user of such methods should be aware of their difficulties and limitations. The loss of data is very high when utilizing EEG techniques, relative to other forms of workload assessment.

MENTAL WORKLOAD MEASUREMENT: CONCLUSIONS

It is the contention of most researchers within the field of mental workload today that the phenomenon is multidimensional (DCIEM, 1988; Eggemeier et al., 1990). Because the majority of workload metrics in use today are unable to assess all dimensions of workload concurrently, a battery of metrics is often necessary to glean as much information as possible. Additionally, certain metrics are stronger with respect to certain criteria, but weaker in other respects. For example, a metric may be highly sensitive to a variety of workload situations, but not very diagnostic of the underlying factors involved. In view of all these factors, DCIEM (1988) has suggested that in order to maximize workload assessment:

> more diagnostic measures should be employed in tandem with more sensitive measures. Techniques that are sensitive to overloads should be used with those that can discriminate workload levels below threshold. Objective measures should be combined with subjective estimates. Workload parameters which represent each component -- task inputs, operator-centered variables, and response outputs -- should be present in an evaluation of mental load. Finally, methods that are sensitive to various operator behaviors should be employed together in multi-task environments (p. 29).

SITUATIONAL AWARENESS

As with workload, there are a number of variations in defining situational awareness (SA) that are accepted in the literature. Some definitions focus on the information processing domain of SA. Examples include those summarized in Garland, Tilden, Blanchard, and Wise (199) and Endsley (1990; 1991a; 1991b). Garland et al. (1991) list several definitions of SA which relate to internal mental models and working memory. Endsley (1990; 1991a; 1991b) provides a three-level definition of SA. "Level I" involves the "perception of the elements in the environment" (Endsley, 1991a, p. 7), "Level II" involves the "comprehension of the current situation" (Endsley, 1991a, p. 7), and "Level III" involves the "projection of future status" (Endsley, 1991a, p. 8). Some definitions are specific to the operational aviation environment. For example, Fracker (1991b), in paraphrasing Sarter and Woods (1991), describes SA as the "military operators' knowledge of the immediate tactical situation" (p. 1). Garland et al. (1991) discuss all aspects of SA solely as they relate to the operational aviation environment. The specificity of these examples and their obvious limitation to the operational and even combat aircraft arena indicate the strong interest in SA associated with this field.

The fact that the term "situational awareness" is so commonplace reflects the high level of interest in this topic. Moreover, the inclusion of aircraft mission terminology in several of the definitions attest to the importance of the concept of situational awareness in the world of aviation.

Adequate SA is considered by aircraft developers to be one of the most important factors in both safety and mission effectiveness (Endsley, 1990; 1991a; 1991b). This is due, in part, to evidence linking a loss in SA as a major contributor to incidents and accidents in both commercial and military aviation. As such, both military aviation officials and the FAA consider the problem of loss of SA one of major concern (Endsley, 1991a; 1991b). According to Endsley (1991a; 1991b) "even the best trained and most experienced pilots can make the wrong decisions if they have incomplete or inaccurate SA" (p. 2)

The link between pilot SA and pilot performance can be understood by first examining the relationship between SA and the internal cognitive or mental model of the aviator.

Internal or mental models consist of learned associations, and are based primarily upon

experience and training (Carmody, 1993). Mental models help to give order to a very complex world, so that the information processor does not have to attend to redundant or irrelevant sources of information. Mental models allow the information processor to recognize patterns of associated events, as well as to acquire and form new associations. Through experience, the human information processor develops a greater repertoire of more and more intricate and representative patterns, enabling him/her to process large amounts of information about a given situation with the utilization of fewer resources. (Braune and Trollip, 1982; Hayes-Roth, 1977; Minsky, 1975). According to Garland et al. (1991),

> "in order to incorporate the vast amounts of data being presented [in today's combat arena], today's pilot must be able to form a dynamic mental representation of the environment surrounding them at all times. This mental representation, in turn, helps to find the causes of observed events, determine the appropriate actions to cause changes, and predict future events. Mental models of the flight environment are seen as fundamental for adequate situation assessment, and resultant situational awareness" (p. 1).

The process of situational awareness can then be stated as maintaining an accurate mental representation of the dynamic external world. In order to accomplish this, the aviator must continually monitor and update the mental model upon which he/she is operating (Carmody and Gluckman, 1993). Barrett and Donnel (1989) refer to such updating as "model transformation". They suggest that "in military systems this process is referred to as the process of maintaining situational assessment" (p. 17).

How critical is this process of maintaining SA to human performance in aviation? In the 1985 symposium for Intraflight Command, Control, and Communications (IFC3), Air Force experts concluded "situational awareness is the single most important factor in improving mission effectiveness" (Garland et al, 1991, p. 3 quoting Beringer and Hancock, 1989). The recognition placed SA as "a critical factor, not only in mission effectiveness, but also in aviation safety, pilot performance, and decision making (Endlsey, 1988c; Garland et al, 1991).

The Measurement of SA

The two most important criteria to be used in evaluating measures of SA are reliability and validity (Fracker, 1991, #128; Garland et al., 1991). Recall that reliability refers to a

metric's ability to provide consistent results over several tests. Validity refers to the metric's ability to measure what it was intended to measure.

Fracker, (1991, #127 and #128) and Garland et al. (1991) consider three types of validity which are important criteria in an SA metric: construct validity, content validity and criterion validity.

Construct Validity

Construct validity addresses the degree to which a metric can assess a psychological construct. A construct is a mental or psychological attribute that is not directly observable (ex: situational awareness), but has the potential to affect observable human behavior.

Assessing construct validity involves the examination of several factors. First, the investigator should have some prior knowledge of what behaviors are related, directly or indirectly, to the construct under examination. Second, the experimenter should have some prior knowledge regarding other constructs that are related to the construct under examination (ex: mental workload and situational awareness). Finally, the experimenter should have some prior knowledge regarding human behaviors related to the other constructs (Fracker, 1991, Garland, 1992).

In assessing the construct validity of SA metrics, there are three recommended criteria (Fracker, 1991, Garland, 1992). First, the measure must distinguish between momentary and reflective situational awareness. In other words, the measure should not confuse the subject's knowledge of what's going on now in the heat of battle with that subject's assessments of the same battle while reflecting upon it hours later in a comfortable setting. Second, the SA measure with consruct validity should indicate a decline in SA when the subject's attention is divided in an increasingly demanding environment. Finally, the SA metric with consruct validity should demonstrate that as situational demands increase, SA should decrease, unless the effort level of the subject increases (Fracker, 1991). A possible exception to this rule is highlighted by Garland et al. (1992). "As pilots become better skilled and experienced, mental effort should decline and SA should decrease." (p. 17). In the latter case, however, the situational demands for the expert have also undoubtedly decreased, due to his/her experience.

Content Validity

Content validity addresses the degree to which an SA metric assesses the appropriate, domain-specific knowledge or behavior of the subjects. In order for this to be accomplished, the experimenter must first analyze the particular situation of interest (ex: tac-air mission scenario) and be familiar with the kinds of information the subject/pilot should know in order to successfully accomplish the task or mission. This information may then be compared to the data gathered by the SA metric, and the latter evaluated for accuracy. Obviously, then, content validity is somewhat specific to the particular mission under investigation. However, Fracker (1988; 1991a; 1991b) outlined five levels of what he terms "situational structure" (Fracker, 1991b, p. 5) that are characteristic of most aviation missions. These levels include goals, organizations, functions, processes, and states.

> In this representation, situations are viewed as sets of variables whose states can change over time. These dynamic variable states are said to result from the interaction of opposing forces each directing their operations toward specific goals. In order to achieve these goals, each force has organized itself into particular units and assigned to each unit specific functions. The interactions among unit functions, referred to as processes, lead directly to the momentary changes in situation variable states (Fracker, 1991b, p. 6).

Utilizing this five-level structure, SA awareness may then be assessed in terms of the degree to which it varies across the five levels. The pilot may exhibit a high level of SA with respect to the enemy plane's objectives, but a low level of SA with respect to the enemy's plan of action. According to Fracker (1991b), an SA metric with high content validity would have the ability to examine subject SA across all five levels.

Fracker's division of SA into levels is similar to Endsley's characterization of SA as a three level process. Recall that Endsley (1990; 1991a; 1991b) also offers a multi-level definition situation awareness as "the perception of the elements in the environment within a volume of time and space [Level I SA], the comprehension of their meaning [Level II SA], and the projection of their status in the near future [Level III SA]" (Endsley, 1991a, p. 7).

Endsley also recommends that "a global measure of SA which simultaneously depicts SA across the many elements of interest is desirable (Endsley, 1991b, p. 8). This recommendation is supported by observations that increases in SA for a particular object or

subtask on a display may accompany unmeasured decreases in SA for other objects or subtasks. This may occur, in part, as a result of subjects biasing their own performance to please the experimenter.

Criterion Validity

Criterion validity refers to "the degree of correlation between the metric and some objective measure that could be used to evaluate the accuracy of a decision based upon the metric" (Fracker, 1991b, p. 6). This criteria is particularly important to the experimenter wishing to demonstrate the relationship between situational awareness and overall mission effectiveness.

Unfortunately, criterion validity is one of the most difficult of the three types of validity to determine. This is because, as highlighted earlier in this document, there are several paths, and several factors within those paths, that can lead to a particular mission or task outcome. Included in these factors are such things as SA, but also decision-making, response efficiency, and perhaps even non-human factors within the system (Endsley, 1990; Fracker et al., 1991b). The ability to establish criterion validity can be aided by using a fairly uniform, highly experienced group of subjects, such as aviators experienced in the mission being assessed. This is because, unlike novices, experienced aviators in a particular situation tend to look for the same types of information (Braune and Trollip, 1982) and, as indicated previously, to achieve relatively consistently high results. On the other hand, with an experienced group, there is the risk of not obtaining statistically significant results in the SA assessment due to a lack of variability (Fracker, 1991b; Garland et al., 1992). In other words, the researcher must carefully plan the mission scenario so as to avoid a ceiling effect. Therefore, it is obvious that, although criterion validity is extremely important, the researcher should be very cognizant of its difficulties and, consequently, very careful in his/her experimental design to assess SA.

SA Measures

There are three main categories of SA: explicit measures, implicit measures, and subjective assessment. A major distinction in SA measures is between explicit and implicit metrics (Fracker, 1991a; 1991b). Explicit measures require subjects to report information from memory, whereas implicit measures are derived from task performance.

Explicit Measures

There are two major categories of explicit measures of SA currently in use. These include retrospective event recall and concurrent memory probes (Fracker, 1991b). Retrospective event recall techniques require subjects to first perform some mission and then recall information about the mission upon completion. Although such techniques have the distinct advantage of very little intrusion in the data collection, there are questions regarding the reliability and validity of retrospective event recall. In the case of reliability, there is simply a lack of evidence. Few studies have examined the reliability of such metrics (Fracker, 1991b). The more serious questions relate to the construct validity of retrospective event recall techniques. There may be a tendency for subjects to produce inaccurate memories, often "filling in the blanks" of what they are unable to actually recall. According to Fracker (1991a), "retrospective recall seems as likely to measure what operators can infer may have happened as what they can remember having actually happened" (p. 2). Similarly, Endsley (1990) considers retrospective event recall, which she classifies as a self-rating subjective technique, highly subject to performance outcomes which may be due to factors other than the subject's SA. In other words, Endsley believes such retrospective reports may reflect more upon subjects confidence levels with respect to their overall performance outcome, as opposed to their level of SA.

Concurrent memory probes offer one solution to the problems of false memories and subject perception. With concurrent memory probes, subjects are queried about relevant information during their performance of a mission (Fracker, 1991a; 1991b). One of the most widely known examples of a concurrent memory probe technique is the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1990; 1991a; 1991b). In Endsley's technique, operational pilots perform a simulated tactical mission. At several random points

in time, the simulation is stopped, the simulation screen is blanked, and the subjects are queried regarding pertinent information present at the time of the stop. As subjects are given questions which have right or wrong answers, percent correct scores determine the pilot's level of situational awareness.

With respect to the reliability of concurrent memory probes, Fracker (1991b) indicates conflicting results in prior studies. Some past research indicated low test-retest reliability (Fracker, 1991a), but Fracker (1991b) suggested these findings might have been due to "idiosyncratic practice effects" (p. 10), as well as the experimenter's measuring SA for the location of objects "with more precision than was psychologically meaningful" (p. 10). Fracker (1991a) did find some indications of reliability in the memory probes for each condition across two experiments. His ultimate conclusion was that further research on the reliability of concurrent memory probes is warranted.

Although in the case of concurrent memory probes, subjects may be less apt to report "false" memories, there still remain questions with respect to validity. Fracker (1991b), for example, questions whether results of concurrent memory probes are subject to working memory decay, thus affecting content validity. If such were the case, later probes (queries more distant in time from the simulation freeze) would actually be more like retrospective event recall, and therefore subject to the same problems. Additionally, Fracker (1991b) suggests that because concurrent memory probes involve sampling the same subjects over several trials on the same mission, the subjects may be prompted to attend to certain information more than others. Both these problems represent potential disadvantages. They may be alleviated through careful planning of the experiment. For example, the investigator may wish only to ask one or two questions at each freeze (thereby restricting possible problems with memory decay). The investigator may also choose not to use the same subjects repeatedly in the same mission, although Fracker (1991b) indicates this approach may be impractical.

With respect to content validity, Fracker (1991b) offers the opinion that concurrent memory probes "can achieve a great deal of content validity, depending upon how they are structured" (p. 12). He indicates that Endsley's (1990) SAGAT has the advantage of high mission specificity and carefully planned technical probes, but he questions the ability of this

technique to sample higher level SA, such as what he would term goal and organization awareness. Recall, however, that Endsley also defines low to high levels of SA, and indicates that SAGAT is able to distinguish among them. The degree of content validity in a concurrent memory probe may then depend upon the SA issues in which the particular investigator is interested. Such SA issues are determined by factors such as the specifics of the mission, what determines successful mission accomplishment, and what questions the experimenter is asking.

Finally, with respect to criterion validity, Fracker (1991b) indicates more research is required to establish correlations between SA and successful mission performance. He does suggest such experiments examine logical links, such as Identification of Friend or Foe (IFF) accuracy (SA measure) and kill probability (mission effectiveness measure).

A potential disadvantage of concurrent memory probes that is not directly related to SA validity is that they are more intrusive than retrospective event recall. If the experimenter wishes to collect other experimental data (such as performance on the simulation task), he/she should consider the possibility of interference from stopping the simulation for the SA queries. However, if only SA data is desired, and the experiment is carefully planned, the concurrent memory probe may prove very useful.

Implicit Measures

With implicit measures of SA, "the goal is to determine whether pilots' mission performance has been influenced appropriately by the occurrence of specific events" (Fracker, 1991b, p. 13). The most widely used examples of implicit measures of SA apply Signal Detection Theory (SDT).

Signal Detection Theory (Green and Swets, 1966) is probably most widely known for its applications in vigilance work. The basic tenets of SDT involve the presentation of "signals" in the context of some background "noise". According to Garland et al. (1992), "a signal detection paradigm is applicable to any situation in which there are two discreet states of the world (signal and noise) that cannot be easily discriminated" (p. 19).

There are four measures of the detection of signals in SDT. These include "hits" (a signal is responded to correctly) and "correct rejections' (the subject does not respond when there is

no signal); "misses" (the subject fails to respond to a signal) and "false alarms" (the subject responds when there is no signal present). A subject's performance on a signal detection task has been shown to be affected by a variety of factors. Among the most substantiated are time-on-task, inter-signal duration and variability, spatial uncertainty, temporal uncertainty, background event rate, signal rate, and subject motivation (Parasuraman, 1976; 1987; Warm, Howe, Fishbein, Dember, and Sprague, 1984; Warm, 1990).

All these factors affect either the sensitivity or the response criterion. The sensitivity factor is the one of primary concern in the measurement of SA, as sensitivity reflects awareness of some phenomenon (Fracker, 1991b). However, Garland et al. (1992) point out that the sensitivity measure typically associated with the signal detection paradigm, d prime (d') may be inadequate for assessments of SA. The sensitivity metric d' indicates the degree of separation between the means of the signal and noise distributions in SDT. Garland et al. (1992) argue that the theoretical curves which describe the noise and signal plus noise distributions, as well as d' and beta (the response criterion) are applicable in tightly controlled laboratory settings, but not so much in realistic and/or very complex situations, such as one would encounter in aviation research. Garland and his colleagues recommend that in such complex and dynamic environments, a better metric for sensitivity is the lesser known A', as it is "not constrained by the rather rigid theoretical assumptions underlying the signal and noise distributions imposed on d'" (Garland et al, 1992 p. 24). This is because A' measures the area under the receiver operating characteristic (ROC) curve that is derived from one data point, as opposed to the noise and the signal plus noise distributions. As such, Garland et al. (1992) suggest that "A' is regarded as the most appropriate sensitivity metric when using an implicit approach to assessing situational awareness in a real-world setting" (p. 24).

The advantages to the use of the signal detection paradigm in assessing SA are that the results are concrete and the immediacy of their recording does not allow for any confusion between momentary and reflective SA (Fracker, 1991b). However, with respect to the reliability of signal detection paradigms in the measurement of SA, more research is necessary, as the present results are somewhat conflicting and do not concretely determine reliability (Fracker, 1991b). The question of validity, on the other hand, again introduces more serious speculation. As is the case with many outcome performance measures, there is

a potential disadvantage to the construct validity, as it is not always easy to separate SA from the other factors possibly affecting the response (decision making, etc.) (Fracker, 1991b). Fortunately, the SDT paradigm, in considering the response criterion, may reduce this to some degree.

With respect to content validity, the potential problems in the SDT paradigm may be more serious. Fracker (1991b) highlights three reasons why the SDT paradigm may present problems in its ability to assess SA for a whole mission segment. First, only single events may be measured by the sensitivity metric. In the case of mission scenarios and simulations, the researcher is typically interested in several simultaneous events. Second, in order to measure a subject's sensitivity to an event using the SDT paradigm, there must be an overt response associated with that event. This may not always be the case, particularly when assessing SA. A researcher may be interested, for example, in the pilot's situational awareness of his plane relative to some other object. Finally, in order to examine false alarms and correct rejections in the SDT paradigm, non-events must be defined. Fracker (1991b) believes this may not be a simple task in non-simulated aviation research.

With respect to criterion validity of the SDT paradigm, Fracker (1991b) reports that no real studies have been conducted. However, it does seem that the SDT paradigm lends itself to logical associations between sensitivity and mission effectiveness (ex: sensitivity to targets and probability of kill). However, similar to the problems with content validity in assessing SA for events not associated with an overt response, the sensitivity metric may not be adequate if the researcher wishes to demonstrate a relationship between such non-response events and mission effectiveness. Again, the questions of the researcher determine, in part, the adequacy of the selected metric.

Subjective Rating Measures

One of the most common forms of SA assessment is the subjective rating metric, due, in part, to its relative ease and non-intrusiveness. There are generally two classes of subjective rating measures: direct and comparative. In the case of direct measures, pilots give an absolute rating to a particular mission previously flown, whereas in the case of comparative measures, this rating is assigned relative to another mission.

Direct Rating Measures

Fracker (1991b) reports that the most common form of subjective rating to assess SA involves Likert scales. These scales have been applied to more global questions of SA (Ward and Hassoun, 1990), as well as on a multi-dimensional level (Arbak, Schwartz, and Kuperman, 1987; Selcon and Taylor, 1989; Taylor, 1989; Venturino, Hamilton, and Dvorchak 1989).

A particular example of a multi-dimensional subjective scaling assessment of SA is the Situational Awareness Rating Technique (SART) (Taylor, 1989). Using SART, pilots rate a new system component or design in terms of three major dimensions: attentional demand, attentional supply, and situational understanding, thereby considering the perceived workload of the pilot in addition to SA (Endsley, 1990; Fracker, 1991, b).

Like the subjective workload scales, subjective assessments of SA have the advantage of being relatively non-intrusive. However, unlike their counterparts in workload research, subjective assessments of SA have not been subject to many reliability studies (Fracker, 1991b).

Endsley (1990) has argued that in the case of SART, because subjects are queried with respect to attentional demand as well as SA, any correlations between performance and the subjective ratings may not necessarily be attributed to SA. This is a general problem with subjective ratings of SA. As indicated by Fracker (1991b), "no coherent theory currently exists either of subjective SA or of how subjective SA might be mapped onto Likert-type rating scales. Consequently, it is difficult to assess just what it is that subjective SA ratings might actually measure." (pp. 17-18). On the other hand, Fracker (1991b) does point out there is some empirical data indicating that SART may have some construct validity, as evidenced by logical correlations between the attentional and SA dimensions of the scale.

With respect to content and criterion validity, there are some difficulties with direct subjective assessments of SA. In the case of content validity, Fracker (1991b) suggests that attempts to balance content and construct validity can be complex, and, for the moment, most research is concerned with establishing a theoretical construct for SA. In the case of criterion validity, on the other hand, the situation is both more serious and, unfortunately, more negative. Much of the available evidence currently shows little correlation between pilots'

subjective ratings of SA and their subsequent mission performance (Venturino et al., 1989; Ward and Hassoun, 1990). Fracker (1991b) has suggested that a possible solution to this problem might be to make pilots aware of the outcomes of the tasks they are performing before giving them the ratings to complete. Fracker feels that much of the inconsistency between subjective assessments of SA and pilot performance may be due to pilots having a higher confidence in their SA than may be warranted. Ultimately, Fracker (1991b) suggests, "subjective SA ratings should not be used alone but should be combined in some way with criterion measures of performance" (p. 19).

Many of the questions and potential problems associated with using subjective rating scales to assess SA are similar to those in using subjective rating scales to assess workload. This is because, to some degree, SA and workload are similar concepts; both involve the "black box" of human information processing. However, there are arguably fewer studies concerning the use of subjective rating scales to assess SA than workload, and this makes the use of the former more challenging.

Comparative Rating Measures

One of the difficulties with direct subjective ratings of SA is that they typically cannot be compared across raters. Additionally, in order to compare within-subjects across conditions, the subjects must be consistent in their use of the rating scale. Fracker (1991b) has indicated that such consistency may be difficult to assess experimentally, and uses this explanation to justify the development of a comparative rating scale technique based on Vidulich's (1989) Subjective WORkload Dominance (SWORD) technique. Recall from the discussion on workload assessment that SWORD asks subjects to make assessments based on the comparison of all possible pairs of tasks performed. In the Fracker and Davis (1990) adaptation of the SWORD technique to SA assessment, subjects perform several experimental conditions, and then compare SA across presented pairs of those conditions. Fracker (1991b) argues that "the fact that subjects directly compare conditions encourages them to apply the same subjective scale to each condition, and the resulting two-way matrix can be examined to determine the extent to which subjects were in fact inconsistent (pp. 19-20).

Once again, a major disadvantage in the application of comparative ratings of SA stems

from the lack of empirical evidence available to support their reliability (Fracker, 1991b). Additionally, many of the problems concerning validity with comparative ratings are the same as with direct ratings of SA, with the exception of content. Because comparative ratings simultaneously examine pairs of conditions, the ratings of SA can be accomplished on several dimensions, thus adding to the content validity of the technique.

SITUATIONAL AWARENESS MEASUREMENT: CONCLUSIONS

Obviously, one of the greatest problems with SA is a lack of empirical data, particularly regarding reliability and certain validity issues. This can only be solved through carefully applied experimentation. Before such research can be undertaken, however, the primary investigator must be very clear as to the definition of SA upon which the empirical inquiries are based. As is notoriously the case with mental workload, there are a number of definitions of SA, not one upon which all have come to full agreement. Unlike workload, however, I believe that there is even less consensus in the conceptual understanding of SA. In other words, although there is no one agreed-upon definition of mental workload, I would venture that most individuals in the field of aviation have roughly the same idea of what it entails. This is not, I would postulate, entirely the case with SA. Although many experts in the field of SA are coming to view it as a multi-level phenomenon, I believe there is still a tendency to focus only upon the perceptual aspects of SA and SA loss. Experimenters should make a very clear argument as to their conceptual definition of SA and the aspects of the phenomenon under investigation before applying a particular methodology.

PHYSICAL METRICS

Perhaps the most straightforward measures, in terms of objectivity, are the physical performance measures. However, as indicated earlier in this document, such measures are not always easily dissected into the components of human performance. The effectiveness of such measures in validating and relating human performance to system performance is dependent on how well the selected measures correspond to the specific research questions.

Classification of Physical Metrics

A review of the relevant literature indicates three general areas of physical metrics in human-aviation systems research: aircraft control metrics, task accuracy metrics, and procedural error metrics.

Aircraft Control Metrics

Aircraft control metrics describe a variety of measures which directly assess the control of the aircraft system. They include purely physical measures, such as the number of stick reversals, as well as physical measures to which some goal-oriented parameters have been attached, such as approach glidescope errors. The measurement provides very direct information about output performance of the human and the total system (depending on the specific measure). Aircraft control metrics provide answers regarding the input and output of human information processing, but not the internal processing.

The following presents a sample list of how aircraft control metrics have been used to assess human performance:

1. The effects of fatigue on pilot performance in pitch, roll and yaw of UH-1H helicopters (Lees, Kimball and Stone, 1977).
2. Army MEDEVAC JUH-1H helicopter pilots were examined for improvements in hovering performance following an applied methodology. Performance metrics included both pilot input measures (cyclic, collective and pedal control inputs) and system output measures (standard deviations in pitch, roll, heading and radar altitude)(Sanders, Burden, Raymond, Simons, Lees, and Kimball, 1978).
3. The effects of the pilots' environment (including ambient temperature, humidity and solar radiation), in combination with the effects of their protective clothing on precision flight and physiological stress was examined. The flight performance metrics included heading, altitude, airspeed and timed turns. All metrics included a predetermined acceptable deviation range (e.g. heading within plus or minus 5 degrees) (Moreland and Barnes, 1969).
4. Low level (LL) and Nap-of-the-Earth (NOE) flight was examined under day conditions and night-with-night vision goggles (NVGs) conditions. Several input and output

measures were used to assess performance in LL and NOE flights. The investigators found that the best discriminators between day and NVG flight for LL were airspeed and the frequency of small control inputs, whereas the best discriminators between visual conditions in NOE flight were the severity of roll angles and the frequency and magnitude of control inputs (Lees, Kimball, Hofman and Stone, 1975).

5. The ability of a Head-Up Display (HUD) design to enhance flight performance under a variety of conditions was examined. Air Force pilots were used as subjects in an F-4C-configured dynamic flight simulator. The aircraft control metrics examined included altitude error (absolute), heading error (absolute), and vertical acceleration (standard deviation) (Soliday and Milligan, 1968).

There are several important factors to consider when selecting particular aircraft control metrics to assess human performance. The researcher must fully understand the aircraft, its functions, capabilities and mission. The responsibilities of the pilot should also be assessed for the particular aircraft and mission via a task analysis. In addition, the research questions will strongly affect the metrics and acceptable parameters that are chosen. The experimenter may also need to examine several aircraft control metrics before determining which are the best discriminators of human and system performance under various conditions. Finally, where aircraft control metrics can provide information regarding what occurred in human performance, they do not, in and of themselves, provide much information as to why. In order to interpret the reasons behind performance enhancements or deficits, metrics which attempt to examine the internal processing of the human may provide additional, invaluable information. This topic will be discussed in detail in a later section.

Task Accuracy Metrics

Task accuracy metrics describe a variety of measures which directly assess the control of a particular task within the aircraft system. As in the case of aircraft control metrics, they include both direct physical measures, such as the number of bombs dropped, as well as those associated with goal-oriented parameters, such as the percentage of targets correctly identified. The latter are typically more common and more useful, as they add the additional dimension

of meaning to the performance metric.

The following presents a sample list of how task accuracy metrics have been used to assess human performance:

1. The performance of helicopter nap-of-the-earth (NOE) reconnaissance aircrew observers in detecting ground targets under various visual and visually aided conditions was examined. The primary task accuracy measure was the number of correct target detections. This metric included the mean number of observers who detected camouflaged targets, mean target detection ranges and mean detection times (Barnes and Doss, 1976).
2. The investigation examined one- versus two-man crews with respect to the performance in visual target detection. Task accuracy measures included the number of targets correctly identified, the number missed, the number falsely identified, and the number of target identification errors.

As is the case with aircraft control metrics, before selecting task accuracy measures, the experimenter should be familiar with the aircraft mission, the pilot/subjects' tasks, and the research questions under consideration.

Procedural Error Metrics

Rather than assessing specific instances or events in time, procedural errors examine sets of behavioral events. These events typically involve either aircraft control or task accuracy metrics, but, in addition, they are evaluated by experts within parameters set by experts. An example of a procedural metric would be the use of one or more expert instructor pilots to rate student pilots in the performance of a mission segment.

The following presents a sample list of how procedural error metrics have been used to assess human performance:

1. The investigation examined the effect of blood alcohol level on the aircraft control and procedural errors of 16 Cessna 172 pilots in ILS approaches. The investigators found increases in approach centerline and glidescope error and variability (which was less pronounced in more experienced pilots) and significant increases in the frequency and seriousness of procedural errors (Billings, Wick, Gerke and Chase, 1973).
2. The investigation examined 80 student pilots in a Vought Air Combat Simulator

configured as an F-4E in a one-versus-one air combat scenario. The analyses included, in addition to procedural error metrics, a carefully selected battery of both aircraft control and task accuracy metrics. The aircraft control and task accuracy measures had been combined into a measure labeled the Good Stick Index (GSI), and included four measures: tracking error, pointing angle advantage (in mean percentage time), ratio of offensive to defensive time, and time to first kill (mean with gun or heat missile). The results indicated that predictions of student pilot performance based on the GSI and on instructor pilot subjective ratings were comparable, and both produced about 75% accuracy in predictions. Additionally, prediction was found to improve to 80% accuracy when both the subjective and objective data were included in the analyses (Moore, Meshier and Coward, 1979).

3. The investigation utilized 23 undergraduate pilots in simulated flight. Both instructor ratings and several aircraft control input and output metrics were examined. Investigators found that in the examination of student performance in straight and level, acceleration, deceleration, climbs, descents, and turns, the aircraft control metrics correlated highly with the instructor ratings and were able to discriminate between pilots of differing experience levels. The investigators further concluded that instructor ratings are a useful method for developing more objective measures of aviator performance (Waag, Eddowes, Fuller, and Fuller, 1975).

4. The investigation examined three groups of 30 pilots in a General Aviation Trainer (GAT) - 2 scenario involving five maneuvers typical of flying under instrument flight rules (IFR). Results indicated that obtaining very high observer-observer reliabilities (r = .771 to .971) is possible if the subjective scale is designed so that performance standards are well defined, descriptions are easy to follow and include details of the maneuver and behavior of interest, and are not too demanding of the rater. Additionally, the investigators found very high correlations (r = .726 to .878) between overt pilot performance and the observers' ratings (Koonce, 1974).

One of the primary concerns when collecting procedural data is, once again, that the experimenter be familiar with the aircraft, its mission, and the pilot's role in that mission. A very important and unique element in the case of procedural metrics is the use of experts to

rate observed behaviors. The advantage of applying one or more experts (> 1 is preferable, as it allows for examination of inter-rater reliability) is that they have inherent knowledge of the task demands, as well a pre-set parameters of acceptance which have been demonstrated to be relatively predictive and accurate. Such expertise can provide a wealth of information in lieu of a detailed and time-consuming task analysis, and can offer insights into student/novice behaviors and thought processes which instrumentation may otherwise be unable to record.

In addition to the three types of physical metrics, there are two general areas presenting issues of concern and interest when applying physical metrics to human performance assessment: the correlation of human state monitoring (ex: eye movements) with physical measurements in order to explain the latter and issues regarding the generalizability of simulator-to-actual flight data.

Employing Human-State Monitoring as a Basis for Interpretation of Physical Performance Metrics

One of the primary benefits to be gained from applying physiological and information-processing human performance data to the interpretation of physical data is the enhanced ability to understand not only what happened, but why. For example, an instructor pilot may be able to record a student's physical performance in handling an aircraft, but if that student performs poorly in that capacity, the instructor will not, on the basis of physical data alone, understand why. Is the student failing to input the appropriate information; failing to concentrate on the correct instruments? Has the student recently experienced a family crisis, which has detrimentally affected his/her concentration? Such factors can severely affect performance outcomes, and in much the same fashion. Corrective actions on the part of the trainer, in this case, are very difficult without understanding the nature of the underlying problem.

The following presents an excellent example of how human state monitoring can be used in conjunction with physical metrics to better understand the underlying causes of human performance outcomes:

The investigation examined 60 Army AH-1 and OH-58 helicopter pilots. The task was to identify specified targets, and the task accuracy measures included number of correct

detections, probability of detections, mean detection time, mean identification range, and mean identification time. In addition, the experiment included the recording of eye movement information. Specifically, the recordings included single glance dwell and fixation times and maximum dwell time. Results indicated large individual differences in target detection times, although the eye fixation data showed only minute differences between the subjects. The combination of this information lead to the interpretation that the large differences in detection time must not be due to the actual sighting of targets, but the underlying decision-making process. In other words, while all subjects fixated upon the targets at about the same time, some subjects had to remain fixated for longer periods of time before reaching the decision that the sighting was, in fact, a target (Barnes, 1978).

Issues Regarding the Generalizability of Data from Simulator to Flight

Obviously, every experimenter must balance the ideal investigative situation with the practical. Although one might like to develop an experimental scenario which virtually duplicates the ultimate environment of interest, issues of economy, safety and practicality often preclude this. Modern simulation technology can very closely match the tactical combat aviation environment, but can never duplicate it. No matter how well the avionics may be matched and the environmental stressors mimicked, the subject (pilot) will always know he is not in any immediate physical danger, and both expectation and motivation can strongly affect performance. This is a particular concern when examining only physical outcome performance data, as the underlying motivations and expectations of the subject are not observed.

The following list presents investigations which highlight some of the concerns when generalizing outcome performance data from simulation to actual flight:

1. The investigation examined 5 Cessna 172 pilots in instrument landing (ILS) approaches in both simulation and actual flight. The aircraft control metrics included approach centerline error (RMSE), approach glidescope error (RMSE) and airspeed deviations (RMSE). Additionally, procedural errors were subjectively assessed and recorded. Results indicated the subjects were more affected by the administration of drugs

in the simulator. Additionally, there was evidence that skills did not necessarily transfer between actual flight and simulation. In other words, experienced and skilled pilots were not necessarily equally skilled in the simulator. The investigators concluded that simulators could be effective in examining stress effects on aviator performance, provided the experimenters realize that the simulator demands a very different strategy for successful flight than in the case of the airplane it is designed to mimic, and that the experimenters understand they must train the subjects to asymptotic performance in the simulator before valid data can be collected (Wickens, Billings, Gerke and Chase, 1974)

2. The investigation examined 22 Navy pilots in a task involving recovery from unusual attitude. Two questions regarding generalizing data from simulation to actual flight were considered. The first concerned the design characteristics of the simulator. Results indicated several things. First, the aerodynamic characteristics of a particular aircraft must be represented as accurately as possible in the simulator, or high correlations between pilots control behaviors in actual and simulated flight will be difficult. Second, the motion characteristics of forward flight in the simulator are very important when large control changes in attitude are required of the pilot. The motion of the simulator is not as important in straight and level flight. Third, providing simulator motion is helpful to the pilot attempting to hover. Finally, the characteristics of the primary visual display in the simulator is important. Large attitude change control maneuvers on the part of the pilot need a wide field of view display.

The second question of generalizing data from simulator to actual flight asked what specific flight events should be considered. Investigators determined the answer to this question very much depended on what information the experimenters wished to derive from the data (Smittle, 1973).

3. The investigation examined three groups of thirty pilots in a General Aviation Trainer (GAT) - 2 in a scenario involving five typical instrument flight rules (IFR) maneuvers. Results indicated the performance of pilots could be predicted from simulator performance, but to differing degrees, depending on the motion characteristics of the simulator. Specifically, the best of three simulator conditions (sustained motion vs washout motion vs no motion) in generalizing from simulator to actual flight performance

was sustained cockpit motion. An advantage to the use of simulation was that measures in the simulator tended to be more reliable across trials (Koonce, 1974)

Despite the potential drawbacks, simulation is typically the safest, most economical and practical solution to data-gathering. It also has the advantage, as with any laboratory situation, of enhanced opportunity for experimental control. Therefore, simulation can prove to be an invaluable method of data collection, as long as the experimenter remains aware of certain critical issues when generalizing from simulation to actual flight, particularly if the latter refers to worst-case scenarios, such as combat.

HUMAN PERFORMANCE MEASUREMENT: CONCLUSIONS

The selection of human performance metrics to be employed in an investigation should always follow a time period in which the experimenters have familiarized themselves with their subjects and subject matter. The importance of understanding the research environment can best be highlighted through an example. Eggemeier et al. (1990) provide such an example of thorough preparation prior to the selection of metrics. Prior to any data collection, they proceeded through a very logical series of steps to arrive at a workload/performance test battery. They first developed a theoretical model to serve as a guide, they became familiar with the mission under investigation and performed a task analysis, and they selected representative subjects (project and operational pilots of the type mission under investigation). Finally, in addition to getting to know their research environment, the investigators got to know their subjects. This was very important, as summarized in the following excerpts, which relate to the development of a subjective procedure by which pilots would describe their allocation of attention:

> "During the interview process, the project pilot SME proposed a method of assessing task prioritization and attention allocation that would be meaningful to pilots........To make this process more understandable to pilots the process was phrased in terms of 'brain bytes'. Brain bytes is a pilot term....." (p. 2-8).

> "At the first step, they were to allocate their attention (100%) to 1 or 3 major classes of functions -- pilot decision making or planning, heads-up activity, and heads-down

activity. This categorization was found to be intuitively appealing since pilots tend to organize their thinking about the cockpit in terms of activities they perform heads-up and heads-down" (p. 2-8).

There are no shortcuts to good experimental methodology. Valid and meaningful experimental results can only be achieved by thorough and contemplative preparation on the part of the researcher(s). Before executing a study in any domain, the investigator must first gain a fairly in-depth understanding of that domain, and/or elicit the help of subject matter experts (SMEs). Particularly when examining a phenomenon as potentially elusive as human information processing and performance in aviation, the background research is critical. The investigator should understand the specific platform he/she is investigating and the specific role of the subject (aviator) in task performance. The investigator must then clearly outline the research questions he/she intends to tackle, and the information desired from the data. Only then can he/she even begin to consider selection of the appropriate human performance metrics.

REFERENCES

Arbak, C.J., Schwartz, N., and Kuperman, G. (1987). *Evaluating the panoramic cockpit controls and displays system*. Paper presented at the 4th Annual Symposium on Aviation Psychology, Columbus, OH.

Anderson, D.B. and Chiou, W.C. (1977). *Physiological parameters associated with extended helicopter flight missions: An assessment of pupillographic data*. (Technical Report No. 77-21), Fort Rucker, Al: U.S. Army Aeromedical Research Laboratory.

Armstrong Aerospace Medical Research Laboratory (AAMRL) (1987). *Subjective workload assessment technique (SWAT): A user's guide* (Draft).

Barnes, J.A. (1978). *A review of individual performance in air-to-ground target detection and identification studies*. Technical Memorandum No. 26-78. Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Barnes, J.A. and Doss, N.W. (1976). *Human engineering laboratory camouflage applications test (HELCAT) observer performance*. (Technical Memorandum No. 7-70), Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Bauer, L., Goldstein, R., and Stern, J. (1987). Effects of information processing demands on physiological response patterns. *Human Factors*, 29, 213-234.

Billings, C.E., Wick, R.L. Jr., Gerke, R.J., and Chase, R.C. (1973). Effects of ethyl alcohol on pilot performance. *Aerospace Medicine*, 44(4), 379-382.

Braune, R.J. and Trollip, S.R. (1982). Towards an internal model in pilot training. *Aviation, Space, and Environmental Medicine*, 53(10), 996-999.

Carmody, M.A. (1993). *Task dependent effects of automation: The role of internal models in performance, workload and situational awareness within a complex, semi-automated cockpit*. Doctoral dissertation, Texas Tech University, Lubbock, TX.

Carmody, M.A., Gluckman, J.P., Morrison, J.G., Hitchcock, E.M., and Warm, J.S. (In progress). *Task-dependent effects of adaptive automation strategies on performance and perceived workload in aviation multi-task scenario*. (Technical Report). Naval Air Warfare Center, Aircraft Division, Warminster, PA.

Casali, J.G. and Wierwille, W.W. (1983). A comparison of rating scale, secondary task,

physiological, and primary task workload estimation techniques in a simulated flight emphasizing communications load. *Human Factors*, 25, pp. 623-641.

Casali, J.G. and Wierwille, W.W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics*, 27, 1033-1050.

Chase, W.G. and Simon, H.A. (1973). The mind's eye in chess. *Visual Information Processing*, 215-281, San Diego: Academic Press, Inc.

Chechile, R.A., Eggleston, R.G., Fleischman, R.N., and Sasseville, A.M. (1989). Modeling the cognitive content of displays. *Human Factors*, 31(1), 31-43.

Comstock, J.R. and Arnegard, R.J. (1990). *Multi-attribute task battery.* (Draft Report), Hampton, VA: NASA Langley Research Center. .

Cooper, G.E. and Harper, R.P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities.* (AGARD Report No. 567), Neuilly-sur-Seine, France: North Atlantic Treaty Organisation.

Davis, D.R. (1948). *Pilot error.* (Air Publication no. 3139A), London England: H.M. Stationary Office.

Defence and Civil Institute of Environmental Medicine (1988). *A preliminary examination of mental workload, Its measurement and prediction.* (Technical Report No. AD-B123-23), Canada: Defence and Civil Institute of Environmental Medicine (DCIEM).

Drew, G.C. (1940). *An experimental study of mental fatigue* (FPRC Report 227). London: Medical Research Council.

Eggemeier, F.T. (1984). Workload metrics for system evaluation. In *Proceedings of the NATO Defence Research Group Panel VIII Workshop on 'Applications of Systems Ergonomics to Weapon System Development'*, pp. C5-C20. Shrivenham, UK: Royal Military College of Science.

Eggemeier, F.T., Biers, D.W., Wickens, C.D., Andre, A.D., Vreuls, D., Billman, E.R., and Schueren, J. (1990). *Performance assessment and workload evaluation systems: Analysis of candidate measures.* (Technical Report No. HSD-TR-90-023), Brooks Air Force Base, TX: Armstrong Aerospace Medical Research Laboratory.

Eggleston, R.G. and Quinn, T.J. (1984). A preliminary evaluation of a projective workload assessment procedure. In *Proceedings of the Human Factor's Society 28th Annual Meeting*, pp. 695-699. Santa Monica, CA: Human Factors Society.

Emery, J.H., Sonneborn, W.G.O., and Elam, C.B. (1967). *A study of the validity of ground-based simulation techniques for the UH-1B helicopter.* (USAAVLAB Technical Report No. 67-72), Fort Eustis, VA: U.S. Army Aviation Material Laboratories.

Endsley, M. R. (1990). *Situation awareness global assessment technique (SAGAT): Air to air tactical version: User's guide.* (NOR DOC no. 89-58 REV A), Hawthorne, CA: Northrop Corporation.

Endsley, M.R. (1991). *Situation awareness in dynamic human decision making: Measurement.* Department of Industrial Engineering, Texas Tech University, Lubbock, TX. Unpublished document.

Endsley, M.R. (1991). *Situation awareness in dynamic human decision making: Theory.* Department of Industrial Engineering, Texas Tech University, Lubbock, TX. Unpublished document.

Fitts, P.M and Jones, R.E. (1947). *Analysis of 270 "pilot error" experiences in reading and interpreting aircraft instruments* (Report TSEAA-694-12A). Wright-Patterson Air Force Base, OH: Aeromedical Laboratory.

Fracker, M.L. (1991). *Measures of situation awareness: An experimental evaluation.* (Report No. AL-TR-1991-0127), Wright-Patterson Air Force Base, OH: U.S. Air Force Aerospace Medical Research Laboratory,

Fracker, M.L. (1991). *Measures of situation awareness: Review and future directions.* (Report No. AL-TR-1991-0128), Wright-Patterson Air Force Base, OH: U.S. Air Force Aerospace Medical Research Laboratory.

Fracker, M.L. and Davis, S.A. (1991). *Explicit, implicit, and subjective rating measures of situation awareness in a monitoring task.* (Report No. AL-TR-1991-0091), Wright-Patterson Air Force Base, OH: U.S. Air Force Aerospace Medical Research Laboratory.

Garland, D.J., Phillips, J.N., Tilden, D.S., and Wise, J.A. (1991). *Theoretical underpinnings of situational awareness: Towards an objective measure.* Center for Aviation/Aerospace

Research, Embry-Riddle Aeronautical University, (Final Technical Report No. CAAR-15498-91-1), for McDonnell Aircraft Company, McDonnel Douglas Corporation, St. Louis, MO.

Garland, D.J., Tilden, D.S., Blanchard, J.W., and Wise, J.A. (1992). *An implicit approach to assessing situational awareness: Development of a situational awareness sensitivity metric.* Center for Aviation/Aerospace Research, Embry-Riddle Aeronautical University, (Final Technical Report No. CAAR-15411-92-2), for McDonnell Aircraft Company, McDonnel Douglas Corporation, St. Louis, MO.

Gillingham, K.K. and Wolfe, J.W. (1986). *Spatial orientation in flight.* (Technical Report No. USAFSAMTR-85-31), Brooks Air Force Base, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

Glenn, F., Cohen, D., Wherry, R., Carmody, M. and Boardway, J. (1993). *Development and validation of a workload assessment technique for cockpit function allocation.* (CHI Systems Inc. Technical Report No. 930730.9000D8), for the Air Vehicle and Crew Systems Technology Department, Human Factors Branch, Naval Air Warfare Center, Aircraft Division, Warminster, PA.

Green, D.M. and Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshtaki (Eds), *Human mental workload* (pp. 139-183). Amsterdam: North-Holland.

Hayes-Roth, B. (1977). Evolution of cognitive structures and processes. *Psychological Review*, 84(3), 260-278.

Iampietro, P.F., Melton, C.E. Jr., Higgins, E.A., Vaughan, J.A., Hoffman, S.M., Funkhouser, G.E., and Saldivar, J.T. (1972). High temperature and performance in a flight task simulator. *Aerospace Medicine*, 43(11), 1215-1218.

Johanssen, G., Moray, N.P., Pew, R., Rasmussen, J., Sanders, A., and Wickens, C.D. (1979). Final Report of the experimental psychology group. In N.P. Moray's (Ed.), *Mental workload: Its theory and measurement*, pp. 101-114, New York: Plenum Press.

Kramer, A.F. (1989). *Physiological metrics of mental workload: A review of recent progress*. University of Illinois, Champaign, Ill. Unpublished manuscript.

Krantz, D.H. and Tversky, A. (1971). Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 79, 151-169.

Koonce, J.M. (1974). *Effects of ground-based aircraft simulator motion conditions upon prediction of pilot proficiency. Parts I and II.* (Technical Report No. AFOSR-TR-74-1292), Arlington, VA: U.S. Air Force Office of Scientific Research.

Krendal, E.S. and Bloom, J.W. (1963). *The influence of selected factors on shrinkage and overfit in multiple correlation*. (NAMI Monograph No. 17), Pensacola, FL: Naval Aerospace Medical Institute.

Lane, N.E. (1986). *Issues in performance measurement for military aviation with applications to air combat maneuvering*. (Technical Report No. NTSC TR-86-008), Orlando, Fl: Naval Training Systems Center.

Lees, M. A., Kimball, K. A., and Stone, L. W. (1977). The assessment of rotary wing aviator precision performance during extended helicopter flights. *Proceedings of the 21st Annual Meeting of the Human Factors Society*, pp. 426-430.

Lees, M.A., Kimball, K.A., Kent, A., Hofmann, M.A., and Stone, L.W. (1975). *Aviator performance during day and night terrain flight*. (USAARL Report No. 77-3), Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Lindholm, E. and Cheatham, C.M. (1983). Autonomic activity and workload during learning of a simulated aircraft carrier landing task. *Aviation, Space, and Environmental Medicine*, 54, 435-439.

Lindholm, E., Sisson, N., Miller, M.J., and Toldy, M.E. (1987). *Physiological assessment of pilot workload in the A-7 aircraft*. (Technical Report No. AD-A178-937), Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

McCracken, J.H. and Aldrich, T.B. (1984). *Analysis of selected LHX mission functions: implications for operator workload and system automation goals*. (Technical Note ASI1479-024-84), Fort Rucker, AL: Anacapa Sciences, Inc.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (ed.), *The*

psychology of computer vision (pp 211-277). New York. McGraw Hill.

Moore, S.B., Meshier, C.W., and Coward, R.E. (1979). The good stick index. A performance measurement for air combat training. *First Interservice/Industry Training Equipment Conference*. Orlando, FL: Naval Training Equipment Center.

Moreland, S. and Barnes, J.A. (1969). Exploratory study of pilot performance during high ambient temperatures/humidity. In *Proceedings, Annual AGARD Symposium for Measurement of Aircrew Performance*, pp. 12-1 - 12-24. Paris: AGARD.

Norman, D.A. and Bobrow, D.G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44-64.

Parasuraman, R. (1976). *Task Classification and decision processes in monitoring behavior.* Unpublished doctoral dissertation, University of Aston, Birmingham.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors*, 29(6), 695-706.

Pepermans, R.G. and Corlett, E.N. (1983). Cross-modality matching as a subjective assessment technique. *Applied Ergonomics*, 14, 169-176.

Rasmussen, J. (1986). *Information processing and human-machine interaction*. Amsterdam: Elsevier North-Holland.

Reid, G.B., Shingledecker, C.A. and Eggemeier, F.T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the IEEE International Conference on 'Cybernetics and Society'*. New York: IEEE.

Reid, G.B., Shingledecker, C.A., Hockenberger, R.L., and Quinn, T.J. (1984). A projective application of the subjective workload assessment technique. In *Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON)*, pp. 824-826. New York: IEEE.

Rolfe, J.M. (1971). The secondary task as a measure of mental workload. In W.T. Singleton, J.G. Fox, and D. Whitfield (Eds.), *Measurement of man at work*, pp. 135-148. London: Taylor and Francis Ltd.

Sander, M.G., Burden, R.T. Jr., Simmons, R.R., Lees, M.A., and Kimball, R.A. (1978). *An Evaluation of perceptual-motor workload during a helicopter hover maneuver.* (USAARL Report No. 78-14), Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Sanders, M.S. and McCormick, E.J. (1987). *Human factors in engineering and design.* New York: McGraw-Hill.

Sanders, M.G., Simmons, R.R., and Hofmann, M.A. (1979). Visual workload of the copilot/navigator during terrain flight. *Human Factors*, 21(3), 369-383.

Sarter, N.B. and Woods, D.D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1, 45-57.

Schneider, W. and Shiffrin, R.M. (1977). Control and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84, 1-66.

Selcon, S.J. and Taylor, R.M. (1989). Evaluation of SART as a tool for aircrew systems design. In *Proceedings of the NATO AGARD Conference on Situational Awareness in Aerospace Operations (AGARD-CP-478).* Springfield, VA: National Technical Information Service.

Shingledecker, C.A., Crabtree, M.S., Simons, J.C., Courtright, J.F., and O'Donnell, R.D. (1980). *Subsidiary radio communications tasks for workload assessment in R&D simulations: I. Task development and workload scaling.* (Report No. AFAMRL-TR-80-126), Wright-Patterson Air Force Base, OH: U.S. Air Force Aerospace Medical Research Laboratory.

Shingledecker, C.A. and Crabtree, M.S. (1982). *Subsidiary radio communications tasks for workload assessment in R&D simulations: II. Task sensitivity evaluation.* (Report No. AFAMRL-TR-82), Wright-Patterson Air Force Base, OH: U.S. Air Force Medical Research Laboratory.

Sirevaag, E., Kramer, A., de Jong, R., and Mecklinger, A. (1988). A psychophysiological analysis of multitask processing demands. *Psychophysiology*, 25, 482.

Skipper, J.H., Rieger, C.A. and Wierwille, W.W. (1986). Evaluation of decision-tree rating scales for mental workload estimation. *Ergonomics*, 29, 585-599.

Soliday, S.M. and Milligan, J.R. (1968). Terrain-following with a head-up display. *Human Factors*, 10(2), 117-126.

Soliday, S.M. and Schohan, B. (1964). *A simulator investigation of pilot performance during extended periods of low-altitude, high-speed flight.* (NASA Report No. CR-63),

Washington, D.C.: National Aeronautical and Space Administration.

Stern, J. and Skelly, J. (1984). The eyeblink and workload considerations. *Proceedings of the Human Factors Society 28th Annual Meeting*, pp. 942-944.

Sternman, M.B., Schummer, G.J., Dushanko, T.W., and Smith, J.C. (1987). Electroencephalographic correlates of pilot performance: simulation and in-flight studies. In *Electric and magnetic activity of the central nervous system: Research and clinical applications in aerospace medicine*. Trondheim, Norway: AGARD Symposium.

Stevens, A. A. (1962). The surprising simplicity of sensory metrics. *American Psychologist*, 17, 29-39.

Taylor, R.M. (1989). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In P*roceedings of the NATO AGARD Conference on Situational Awareness in Aerospace Operations (AGARD-CP-478)*. Springfield, VA: National Technical Information Service.

Thiessen, M.S., Lay, J.E., and Stern, J.A. (1986). *Neuropsychological workload battery validation study*. (General Dynamics Report No. FZM 7446), Fort Worth, TX., for Harry G. Armstrong Aerospace Medical Research Laboratory.

Thorton, D.C. (1985). An investigation of the Von Restorff phenomenon in post-test workload ratings. In *Proceedings of the Human Factors Society 29th Annual Meeting*, pp. 760-764. Santa Monica, Ca: Human Factors Society.

Venturino, M., Hamilton, W.L., and Dvorchak, S.R. (1989) Performance-based measures of merit for tactical situation awareness. *Proceedings of the NATO AGARD Conference on Situational Awareness in Aerospace Operations (AGARD-CP-478)*. Springfield, VA: National Technical Information Service.

Vidulich, M.A. (1989). The use of judgment matrices in subjective workload assessment: The Subjective WORkload Dominance (SWORD) technique. *Proceedings of the Human Factors Society 33rd Annual Meeting*, pp. 1406-1410.

Vidulich, M.A., Ward, G.F., and Schueren, J. (1991). Using the Subjective WORkload Dominance (SWORD) technique for projective workload assessment. *Human Factors*, 33(6), 677-691.

Waag, W.L., Eddowes, E.E., Fuller, J.H. Jr., and Fuller, R.R. (1975). *ASUPT automated objective performance measurement system.* (Technical Report No. AFHRL-TRL-75-3), Williams Air Force Base, AZ: Air Force Human Resources Laboratory.

Ward, G.F. and Hassoun, J.A. (1990). *The effects of head-up display pitch ladder articulation, pitch number location and horizon line length on unusual attitude recoveries for the F-16.* (Technical Report No. ASD-TR-90-50008), Wright-Patterson AFB, OH: Aeronautical Systems Division.

Warm, J.S. (1990). Vigilance and target detection. In C.D. Wickens and B. Huey (Eds), *Teams in transition: Workload, stress, and human factors.* Washington, D.C.: National Research Council.

Warm, J.S., Howe, S.R., Fishbein, H.D., Dember, W.N. and Sprague, R.L. (1984). Cognitive demand and the vigilance decrement. In A. Mital (Ed.) *Trends in ergonomics/human factors I.* Elsevier Science Publishers B.V., North-Holland.

Wick, R.L., Billings, C.E., Gerke, R. J., and Chase, R.C. (1974). *Aircraft-simulator transfer problems.* (Technical Report No. ARMRL-TR-74-68), Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratory.

Wickens, C.D. (1984). *Engineering psychology and human performance.* Columbus, OH: Merrill.

Wiener, C.D. and Nagel, D.C. (Eds.). (1988). *Human factors in aviation*, San Diego, CA: Academic Press, Inc.

Williges, R.C. and Wierwille, W.W. (1979). Behavioral measures of aircrew mental workload. *Human Factors*, 21, 549-574.

Wierwille, W.W. and Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications. *Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting*, pp. 129-133.

Wierwille, W.W., Rahimi, M. and Casali, J.G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors*, 27, 489-502.

Wierwille, W.W. and Williges, R.C. (1978). *Survey and analysis of operator workload*

assessment techniques. (Technical Report No. S-78-101), Patuxent River, Maryland: Naval Air Test Center.

Williams, A.C. (1947). Preliminary analysis of information required by pilots for instrument flight. In S.N. Roscoe (Ed.), *Aviation Research Monographs*, 1(1). Urbana: University of Illinois, Aviation Research Laboratory.

Wilson, G.F. (1991). *Progress in the psychophysiological assessment of workload*. (Technical Report No. AL-TR-1992-0007), Wright-Patterson Air Force Base, OH: Armstrong Laboratory.

Wilson, G.F., O'Donnell, R.D., and Wilson, L. (1982). *Neurophysiological measures of A-10 workload during simulated low altitude missions*. (Report No. AFAMRL-TR-83-0003), Wright-Patterson Air Force Base, OH: U.S. Air Force Aerospace Medical Research Laboratory.

Wilson, G.F., Purvis, B., Skelly, J., Fullenkamp, P., and Davis, I. (1987). Physiological data used to measure pilot workload in actual flight and simulator conditions. *Proceedings of the Human Factors Society 31st Annual Meeting*, 779-783.

DISTRIBUTION LIST (Continued)

	No. of Copies
Defense Technical Information Center ATTN: DTIC-FDAB Cameron Station BG5 Alexandria, VA 22304-6145	2
Center of Naval Analysis 4401 Fort Avenue P.O. Box 16268 Alexandria, VA 22302-0268	1
1299th Physiological Training Flight Malcolm Grow USAF Medical Center Andrews AFB, Washington, DC 20331-5300	1

DISTRIBUTION LIST

	No. of Copies
Dr. Jim Ballas Naval Research Lab Code 5534 4555 Overlook Avenue, SW Washington, DC 20375-5000	1
LCDR John Deaton Defense Training and Performance Data Center 3280 Progress Dr. Orlando, FL 32826-3229	10
Dr. Peter Hancock 172 Pillsbury Dr. SE 164 Norris Hall Minneapolis, MN 55455	1
CDR S.D. Harris Naval Air Systems Command Code AIR-05TP3 Washington, DC 20361-3300	1
CDR J.M. Owens Naval Research Lab Code 5510 4555 Overlook Avenue, SW Washington, DC 20375-5000	1
Dr. Raja Parasuraman Dept. of Psychology Catholic University of America Washington, DC 20064	1
Dr. John Reising Cockpit Integration Directorate Code WRDC/KTC Wright-Patterson AFB, OH 45433	1
Dr. Joel Warm University of Cincinnati Dept. Of Psychology Cincinnati, OH 45221	1
Naval Air Warfare Center - Aircraft Division (2 for Code 8131)	2